Exome array analysis of pulmonary function in smokers with and without chronic obstructive pulmonary disease (COPD)

By Margaret M. Parker, MHS

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

June, 2015

Abstract:

Chronic obstructive pulmonary disease (COPD) is a progressive respiratory disease characterized by airflow restriction and decreased lung function. It is the 3rd leading cause of death worldwide, accounting for approximately 3 million deaths in 2010. It is diagnosed using spirometric measurements, including the forced expiratory volume in one second (FEV_1) and the forced vital capacity (FVC). These measures reflect the severity of airway obstruction and predict population morbidity and mortality. The primary environmental cause of COPD is cigarette smoking, but genetics also play a role in individual susceptibility and disease progression. Genome-wide association studies (GWAS) have identified over 30 loci associated with lung, but together the identified variants can only explain a small proportion of the variation in spirometric measures and a small proportion of the estimated heritability. We hypothesized: 1) rare functional variation also affects lung function; and 2) genetic variation (both common and rare) affects longitudinal changes in lung function. This dissertation tests these hypotheses using data from the COPDgene study, a large multicenter study of current and former smokers.

Thesis committee

Terri Beaty, Ph.D. (advisor)
Professor of Epidemiology

Ingo Ruczinski, Ph.D., M.S.
Professor of Biostatistics

Rasika Mathias, Sc.D.
Associate Professor of Medicine

Thesis readers

Terri Beaty, Ph.D. (advisor)
Professor of Epidemiology

Priya Duggal, Ph.D.
Associate Professor of Epidemiology

Rasika Mathias, Sc.D.
Associate Professor of Medicine

Margaret Taub, Ph.D.
Assistant Scientist of Biostatistics

Acknowledgements

My dissertation would not be possible without the help of many people. First and foremost, I would like to thank my advisor Dr. Terri Beaty for all her guidance and support. Thanks for encouraging me to get involved in many different projects and thanks for all the opportunities you have given me. I would not have made it through the PhD process without you.

I would also like to thank my thesis committee, Dr. Ingo Ruczinski and Dr. Rasika Mathias. Thank you for all the time you spent in my committee meetings, thanks for sharing your expertise with me, and thanks for the encouragement along the way. I truly appreciate it.

This dissertation would not be possible without the help of my COPDgene collaborators, especially Michael Cho, Brian Hobbs, and Ed Silverman. Thanks for the opportunity to work on this project, and thanks for all of your input along the way.

I sincerely thank all the faculty and students in the genetic epidemiology department. I have always felt supported by the whole department and was lucky to have many "advisors" along the way. A special thanks to Dr. Priya Duggal, Dr. Linda Kao, and Dr. Dani Fallin and for teaching me genetic epidemiology, for pushing me to do a PhD, and for mentoring me along the way. Your encouragement and enthusiasm meant a lot to me.

Thanks to everyone in the Beaty lab group. You guys provided a great environment and supported me through the dissertation process. Thanks to all those that sat in W6517 with me at some point including: Gen, Pooj, Stephanie, Tao, Tanda, and Jackie. You guys have taught me so much and been like a second family. I would not have made it

through this process without your support, advice, and friendship. Thanks for keeping me laughing.

I would like to thank my wonderful friends. You made grad school fun and were always there to help me celebrate the highs and survive the lows. A PhD is a group effort and you guys were the best group I could ask for.

Lastly, I would like to thank my family for their continuing support. To my parents, Jane and Tony, and to my siblings, Beebs and Scott, I could not have done it without you!

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction and literature review

Chapter 1. Introduction and literature review

Chronic obstructive pulmonary disease (COPD) is a progressive respiratory disease characterized by airflow restriction and decreased lung function. COPD results from both small airway disease (obstructive bronchitis) and parenchymal destruction (emphysema) and typical symptoms include dyspnea, chronic cough, and chronic sputum production. These symptoms tend to worsen over time, as and the pathological changes associated with COPD are not fully reversible. Individuals with COPD can vary greatly in symptom severity, but disease progression typically leads to a decreased capacity for exercise, an increased risk of serious comorbidites, and an overall increased mortality rate[1,2].

COPD burden

COPD is a highly prevalent disease resulting in a substantial economic burden that is increasing[3]. According to the WHO's Global Burden of Disease, Injuries, and Risk Factors Study, COPD is the 3rd leading cause of death worldwide, accounting for approximately 3 million deaths in 2010[4]. It is estimated 4.2% of working adults in the United States (or 15 million individuals) have COPD, and this number is projected increase as the population ages[5]. This not only poses a large medical burden, but also an economic one. The National Heart, Lung and Blood Institute (NHLBI) estimates that along with asthma, COPD costs exceeded $60 billion USD in 2008, of which $53.7 billion USD were direct costs[6]. The high prevalence and massive economic burden of COPD highlight the need for greater understanding of disease etiology to better predict individual risk, prevent future disease and treat diagnosed cases.

Measurement of COPD

The airflow restriction characteristic of COPD is most often quantified via spirometry, including forced expiratory volume in one second (FEV_1), forced vital capacity (FVC),

and the ratio of FEV_1/FVC. These measurements quantify the volume and speed at which an individual can exhale air, providing a useful indication of lung function. To obtain spirometric measurements, patients are asked to take a deep breath and exhale into a sensor as hard as possible, for as long as possible. From this maneuver, one can measure: 1) the volume of air exhaled in one second (FEV_1); 2) the total volume of air exhaled (FVC); and 3) FEV_1 as a percent of the predicted value based on height, age, sex, and race (FEV_1 percent predicted).

Individuals without obstructive airway disease can expire all (or nearly all) of their vital capacity in one second, resulting in an FEV_1/FVC ratio about equal to one. Alternatively, an FEV_1/FVC ratio below 0.7 defines airflow obstruction and is used to diagnose COPD (Figure 1.1).

Spirometric measurements are easily obtained, inexpensive and provide useful benchmarks for COPD progression. Moreover, they reliably predict population mortality. In a longitudinal study of 15,759 individuals followed over 11 years, those with the most severe lung function impairment (FEV_1/FVC < 0.7 and FEV_1 % predicted < 50%) were at a 5.7 times higher risk of death (95% CI = 4.4 - 7.3) than those with normal lung function, highlighting the importance of these measurements as indicators of population health[7].

However, spirometry is not without limitations. The measurement procedure is highly dependent on patient cooperation and effort, and therefore can vary substantially among individuals[8]. Additionally, spirometry does not fully describe all aspects of disease pathogenesis, including changes in airway thickness, emphysema and severity of symptoms. Thus, while these measurements can confirm the presence of airflow obstruction, they do not provide any specific etiologic diagnosis[3,9].

Etiology of COPD

COPD is a multifactorial disease with the primary cause being cigarette smoking[10]. However, only 25% of cigarette smokers ever develop COPD[11] and even among those that do, the rate of lung function decline varies considerably among smokers with similar exposure levels[12,13]. This suggests additional risk factors, including genetics, may play some role in individual susceptibility to COPD development and disease progression.

Environmental risk factors

Cigarette smoking

The primary environmental cause of COPD is cigarette smoking. In a 25-year prospective cohort study of 8,045 individuals, current smokers were 6.3 times more likely to develop clinically significant COPD than never smokers (95% CI = 4.2-9.5)[4]. Approximately 80% of COPD deaths are caused by smoking, and smokers are 12-13 times more likely to die from COPD than never smokers[14].

Cigarette smoking leads to airflow obstruction and COPD symptoms through a complex physiological process. Repeated exposure to the irritants in cigarette smoke causes chronic inflammation of the airways, leading to an influx of inflammatory mediator molecules (e.g. neutrophils, B cells, and T-lymphocytes)[15]. Over time, this chronic inflammation causes irreversible structural and physiological changes in the lung including airway constriction, excess mucus production and dysfunctional cilia[16]. These changes result in expiratory airflow limitation, especially in the smaller (< 2mm) airways, and hallmark COPD symptoms, such a wheezing, coughing and dyspnea[17].

Sex

There are distinct sex differences in COPD prevalence. Overall, men are more likely to have COPD, but over the past 20 years COPD prevalence and mortality have increased

more rapidly among women[18]. Sex differences in COPD are frequently attributed to different exposure rates (e.g. smoking) between men and women, but there is increasing evidence that biological factors may also play a role. Female participants of the National Emphysema Treatment Trial (NETT) had fewer pack-years of smoking history than men but had similarly severe COPD[19]. Additionally, a meta-analysis by Gan et al. found that females had a faster annual rate of decline in lung function (measured as FEV_1) than males of similar smoking levels[20]. Overall, research suggests women develop more severe COPD at younger ages with lower levels of exposure to smoking[19–21]. Proposed explanations for this disparity include: 1) differential susceptibility to the effects of tobacco, 2) anatomical differences (e.g. smaller lungs and airways), and 3) differential responses to treatments[22]. The effect of sex on COPD development and progression is likely due to a combination of environmental and biological factors, however the exact mechanism resulting in this disease disparity remains unresolved.

Race

Differences in lung function between European Americans (EAs) and African Americans (AAs) are well documented. Prevalence of any impairment in lung function is higher among EAs than AAs[23]. According to the CDC, the age adjusted mortality rate from COPD is 46.0 per 1000 in EA and 27.2 per 1000 in AA individuals[24].

However, mortality from COPD is increasing more rapidly among AAs[25], and there is evidence that AAs may be more susceptible to the negative effects of tobacco smoke[26,27]. In 2004, Chatalia et al. reported that among 80 EA and 80 AA patients with advanced COPD, AAs were younger with less cumulative smoking history despite comparable lung function[28]. More recently, differences in between EA and AAs have been studied in the COPDgene study, an observational case-control study of 10,000 smokers with and without COPD[29]. Investigators found AAs were much more likely to

have severe-early onset COPD than EAs (42% vs. 14%) and had different patterns of emphysema distribution over the lung region[30]. The potential impact of racial differences in COPD development has important implications on disease is screening, diagnosis and treatment, and more research is needed to understand why racial disparities may exist.

Age

Aging is associated with a progressive decline in lung function even among non-smoking adults[31]. It remains unclear if this change results from the normal aging process (i.e. age-related changes in pulmonary mechanics, respiratory muscle strength, gas exchange and ventilatory control)[32] or if age reflects the sum of cumulative exposures throughout life leading to lung function decline[3].

Other environmental risk factors

A number of additional risk factors are associated with an increased risk of COPD including occupational dust exposure[21,34], air pollutants[35], history of respiratory infection [36], asthma[37–39], nutrition[40,41], second-hand smoke exposure[42], and socio-economic status[43]. The relative contribution of these diverse risk factors to COPD development and progression remains unresolved, and many may be especially important contributors to disease in non-smokers[3].

Genetic risk factors

There is substantial evidence that genetic factors can influence COPD susceptibility[44–51]. Early familial aggregation studies found severe, early onset COPD probands were significantly more likely to have first-degree relatives with decreased FEV_1 than controls[51]. More recently, a large population-based study of 821 cases and 776 smoking controls showed individuals with a parental family history of COPD were 1.73 times more

likely to have disease than those without a family history[49], indicating COPD aggregates in families which could reflect either shared genetics or shared environment.

The estimated heritability, or proportion of variance in lung function attributable to the additive effects of genes, is 52%, 54% and 45% for FEV_1, FVC and FEV_1/FVC (respectively). This means a considerable proportion of the variation in these spirometric measurements is due to genetic factors[50].

Alpha-1 antitrypsin deficiency

The best understood genetic risk factor of COPD is a mutation on chromosome 14 in the gene coding the serine protease inhibitor alpha-1 antitrypsin. The role of alpha-1 antitrypsin deficiency in lung disease was first identified in 1963 by Carl-Bertil Laurell when he noted the absence of alpha-1 antitrypsin protein in the plasma of many early onset emphysema cases[52]. This absence was eventually traced back to a missense mutation in the alpha-1 antitrypsin gene that occurs when a glutamic acid is substituted for a lysine at amino acid position 342 leading to a non-functional protease inhibitor[52]. Individuals homozygous for this null mutation display markedly decreased alpha-1 antitrypsin levels, typically leading to lung tissue degradation and early-onset, severe COPD[53]. Although well-characterized, this mutation is relatively rare in the population, and accounts for only 1-2% of all COPD cases, suggesting other genetic factors control disease susceptibility[54].

Genome-wide association studies

A number of investigators have attempted to characterize common genetic variation influencing lung function using genome-wide association studies (GWAS). In 2009, Pillai et al. identified an association between variants in the *CHRNA3/CHRNA5* gene cluster on chromosome 15q25 and COPD case-control status in 823 cases and 810 smoking

controls of European descent[55]. Subsequent work has replicated this association and implicated a nearby gene in tight linkage disequilibrium, *IREB2*, as the likely candidate gene underlying this association[56]. Concurrent to the Pillai et al. GWAS, Wilk et al. conducted an independent GWAS of lung function in 7,691 Framingham Heart Study participants, using FEV_1/FVC ratio as a quantitative outcome[57]. They identified 4 single nucleotide polymorphisms (SNPs) near the hedgehog interacting protein (*HHIP*) gene as genome-wide significant ($p < 5x10^{-8}$). *HHIP* is a regulatory protein in the hedgehog signaling pathway and may influence fetal lung development, although its role in lung disease remains unclear[57]. Finally, Cho et al. identified an association between *FAM13A* loci and case-control status using in 2,940 cases and 1,380 controls from 3 independent study populations[58] (ECLIPSE[59], NETT[60], and Norway[61]). In summary, early GWASs with modest sample sizes identified three genes (*IREB2*, *HHIP*, and *FAM13A*) associated with COPD and its related qualitative outcomes.

Large scale meta-analyses

More recently, investigators have leveraged the large sample sizes afforded by ongoing cohort studies to test for association between common markers and lung function outcomes. Three large meta-analyses of GWAS data have identified 25 additional loci associated with lung function outcomes. In 2010, Hancock et al. identified 4 novel loci associated with FEV_1/FVC in 20,890 participants of European descent in the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) studies. Associated SNPs were located in or near the genes: *GPR126, ADAM19, PTCH1 and PID1*[62]. Concurrently, Repapi et al. identified 5 novel loci in 20,288 individuals who were part of the SpiraMeta Consortium studies. Associated markers were located in the following genes: *TNS1, GSTCD, HTR4, AGER, THSD4*[63]. Finally, in the largest meta-analyses of lung function to date, Soler-Artigas et al. combined these two large meta-analyses

(CHARGE and SpiraMeta) to evaluate 2.5 million SNPs in 48,201 individuals of European ancestry. This analysis identified 16 novel genetic loci associated with FEV_1/FVC including variants in and near the genes: *MFAP2, TGFB2, HDAC4, RARB, MECOM, SPATA9, ARMC2, NCR3, ZKSCAN3, CCDC38, C10orf11, LRP1, CCDC38, MMP15, CFDP1 and KCNE2* [64]. Previously identified associations between genetic markers and lung function in European derived populations are summarized in Figure 1.2.

Lung function decline

COPD development is likely influenced by both: 1) impaired attainment of maximal lung size and function before adulthood; and 2) accelerated lung function decline. Cross-sectional studies cannot differentiate between these pathways, and it is probable different risk factors (both genetic and environmental) separately influence each pathway. However, few studies have assessed the genetics of lung function decline, as this requires longitudinal data.

To date, there have been three published GWASs of lung function decline, all in populations of European ancestry. A summary of these studies is provided in Table 1.1. Overall, 6 genes have been identified as associated with lung function decline, none of which overlap the reported associations from studies of cross-sectional lung function[65–67]. However, there is little agreement between the 3 published studies regarding which genes contribute to this decline phenotype, suggesting additional replication and further research is necessary.

Genetic risk factors in African Americans

Despite increasing evidence that African Americans may be especially susceptible to the negative effects of tobacco smoke[26,27], there has been little analysis of the genetic

determinants of lung function in this sub-population. GWAS analysis in 3,260 AA participants of the COPDgene study identified one locus significantly associated with FEV_1/FVC (near *BC011998* on chromosome 5, p-value=$1.31x10^{-8}$) and one marginally significant locus associated with FEV_1 (near *MGAT3* on chromosome 22, p-value = $9.19x10^{-8}$)[68]. Additionally, admixture mapping of this population identified an intronic variant in *FAM19A2* as being associated with FEV_1/FVC[69]. However, compared to European-derived populations, few studies have been conducted in African Americans, and little is known about the genes controlling lung function in this sub-population.

Motivation and specific aims for this study

Substantial advances have been made in our understanding of the genetic etiology of COPD through GWAS and candidate gene studies. However, despite large sample sizes, together all known associated markers account for only 3.2% of the estimated heritability of FEV_1/FVC and 1.5 % of the estimated heritability of FEV_1[70]. Moreover, little is known about the genetic determinants of these measurements in African Americans. Together, this suggests much of the genetic variation controlling reduced pulmonary function has yet to be discovered. We seek to address this issue though the following specific aims:

1. **To identify functional genetic variants associated with lung function in European and African American participants of the COPDgene study using exome array data.**

2. **To identify genetic variants associated with longitudinal changes in lung function in European American and African American participants of the COPDgene study using GWAS and exome array data.**

Study description

COPDgene is an observational study conceived in 2008 to investigate the genetic and environmental etiology of COPD[29]. Participants included 10,280 self-identified AAs or EAs between the ages of 45 and 80 years with a minimum of 10 pack-years of smoking history recruited from 21 study centers across the United States. Study exclusion criteria included those with a history of lung disease (other than asthma), previous surgical removal of at least one lung lobe, active cancer treatment, suspected lung cancer, metal in the chest, exacerbated COPD treated with antibiotics/steroids, known alpha-1 antitrypsin deficiency, recent eye surgery, and inability to use albuterol, self-identified as multiple racial categories, having a first or second degree relative in the study, and pregnant women. Subjects completed detailed questionnaires, pre- and post-bronchodilator spirometry, volumetric computed tomography (CT) of the chest, and provided a DNA sample for genotyping.

Genotyping platforms

Available genotyping platforms included: 1) The Illumina Omni Express BeadChip GWAS array containing approximately 733,000 common single nucleotide polymorphic (SNPs) variants; and 2) Illumina HumanExome chip 12v1-1 (or 12v1-2) exome array containing approximately 233,000 (mostly) functional genetic variants located in gene exons.

Longitudinal follow-up

In 2013, eligible COPDgene participants returned for a 5-year follow-up assessment, affording investigators the opportunity to assess determinates (both genetic and environmental) of decline in lung function. As of November 2015, longitudinal data from 2,000 participants was available for analysis.

Summary

Many recent advances have been made in our understanding of COPD risk factors. However, important determinates (both genetic and environmental) remain unidentified. This dissertation aims to identify functional genetic risk factors associated with reduced lung function and longitudinal lung function decline, using exome and genome-wide array data. Ultimately, improved understanding of COPD risk factors of can aid in tailoring treatment, discovering novel therapeutic interventions and in developing effective prevention strategies, and this is especially timely given the large and increasing global burden of COPD.

References

1. Hole DJ, Watt GC, Davey-Smith G, Hart CL, Gillis CR, Hawthorne VM. Impaired lung function and mortality risk in men and women: findings from the Renfrew and Paisley prospective population study. *BMJ*. 1996;313:711-715; discussion 715-716. doi:10.1136/bmj.313.7059.711.

2. Schunemann HJ. Pulmonary Function Is a Long-term Predictor of Mortality in the General Population: 29-Year Follow-up of the Buffalo Health Study. *CHEST J*. 2000;118:656. doi:10.1378/chest.118.3.656.

3. Global Strategy for Diagnosis, Management, and Prevention of COPD Global Initiative for Chronic Obstructive Pulmonary Disease. 2015. Accessed through http://www.goldcopd.org/uploads/users/files/GOLD_Report_2015.pdf (2/2/2015).

4. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2095-2128. doi:10.1016/S0140-6736(12)61728-0.

5. Doney B, Hnizdo E, Syamlal G, et al. Prevalence of chronic obstructive pulmonary disease among US working adults aged 40 to 70 years. National Health Interview Survey data 2004 to 2011. *J Occup Environ Med*. 2014;56(10):1088-1093. doi:10.1097/JOM.0000000000000232.

6. National Heart, Lung, and Blood Institute . Morbidity and Mortality: 2012 Chart Book on Cardiovascular, Lung and Blood Diseases. *Natl Institutes Heal*. 2012.

7. Mannino DM, Doherty DE, Buist a. S. Global Initiative on Obstructive Lung Disease (GOLD) classification of lung disease and mortality: Findings from the Atherosclerosis

Risk in Communities (ARIC) study. *Respir Med*. 2006;100:115-122. doi:10.1016/j.rmed.2005.03.035.

8. Crapo RO. Pulmonary-function testing. *N Engl J Med*. 1994;331(1):25-30. doi:10.1056/NEJM199407073310107.

9. Barisione G, Brusasco C, Garlaschi A, Crimi E, Brusasco V. Lung function testing in COPD: when everything is not so simple. *Respirol case reports*. 2014;2(4):141-143. doi:10.1002/rcr2.72.

10. Decramer M, Janssens W, Miravitlles M. Chronic obstructive pulmonary disease. *Lancet*. 2012;379(9823):1341-1351. doi:10.1016/S0140-6736(11)60968-9.

11. Løkke A, Lange P, Scharling H, Fabricius P, Vestbo J. Developing COPD: a 25 year follow up study of the general population. *Thorax*. 2006;61(11):935-939. doi:10.1136/thx.2006.062802.

12. Burrows B, Knudson RJ, Cline MG, Lebowitz MD. Quantitative relationships between cigarette smoking and ventilatory function. *Am Rev Respir Dis*. 1977;115(2):195-205. http://www.ncbi.nlm.nih.gov/pubmed/842934. Accessed February 16, 2015.

13. Fletcher CM. Letter: Natural history of chronic bronchitis. *Br Med J*. 1976;1(6025):1592-1593. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1640524&tool=pmcentrez&rendertype=abstract. Accessed February 16, 2015.

14. U .S . Department of Health and Human Services . The Health Consequences of Smoking . A Report of the Surgeon General. 2004.

15. Ali Al Talag Pearce Wilcox, MD, FRCP MD. Clinical physiology of chronic obstructive pulmonary disease. *B C Med J*. 2008;50(March):102.

16. Lapperre TS, Sont JK, van Schadewijk A, et al. Smoking cessation and bronchial epithelial remodelling in COPD: a cross-sectional study. *Respir Res*. 2007;8:85. doi:10.1186/1465-9921-8-85.

17. (ATS) ATS. Standards for the Diagnosis and Management of Patients with COPD. 2004.

18. WHO. World Health Statistics:2008. *Geneva:WHO*. 2008. Available from: http://www.who.int/whosis/whostat/EN_WHS08_Full.pdf. Accessed Dec 11, 2014.

19. Martinez FJ, Curtis JL, Sciurba F, et al. Sex differences in severe pulmonary emphysema. *Am J Respir Crit Care Med*. 2007;176(12):243-252. doi:10.1164/rccm.200606-828OC.

20. Gan WQ, Man SFP, Postma DS, Camp P, Sin DD. Female smokers beyond the perimenopausal period are at increased risk of chronic obstructive pulmonary disease: a systematic review and meta-analysis. *Respir Res*. 2006;7:52. doi:10.1186/1465-9921-7-52.

21. Murray CJ, Lopez AD. Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study. *Lancet*. 1997;349(9064):1498-1504. doi:10.1016/S0140-6736(96)07492-2.

22. Aryal S, Diaz-guzman E, Mannino DM. Influence of sex on chronic obstructive pulmonary disease risk and treatment outcomes. 2014:1145-1154.

23. Halldin CN, Doney BC, Hnizdo E. Changes in prevalence of chronic obstructive pulmonary disease and asthma in the US population and associated risk factors. *Chron Respir Dis*. 2015;12(1):47-60. doi:10.1177/1479972314562409.

24. Centers for Disease Control and Prevention. National Center for Health Statistics.CDC Wonder Online Database. *Ser 20 No 2L*. 2009.

25. Dransfield MT, Bailey WC. COPD: racial disparities in susceptibility, treatment, and outcomes. *Clin Chest Med*. 2006;27(3):463-471, vii. doi:10.1016/j.ccm.2006.04.005.

26. Dransfield MT, Davis JJ, Gerald LB, Bailey WC. Racial and gender differences in susceptibility to tobacco smoke among patients with chronic obstructive pulmonary disease. *Respir Med*. 2006;100(6):1110-1116. doi:10.1016/j.rmed.2005.09.019.

27. Wise RA. Changing smoking patterns and mortality from chronic obstructive pulmonary disease. *Prev Med (Baltim)*. 26(4):418-421. doi:10.1006/pmed.1997.0181.

28. Chatila WM, Wynkoop WA, Vance G, Criner GJ. Smoking patterns in African Americans and whites with advanced COPD. *Chest*. 2004;125(1):15-21. http://www.ncbi.nlm.nih.gov/pubmed/14718415. Accessed March 5, 2015.

29. Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDgene) study design. *Epidemiology*. 2011;7(1):1-10. doi:10.3109/15412550903499522.Genetic.

30. Hansel NN, Washko GR, Foreman MG, et al. Racial differences in CT phenotypes in COPD. *COPD*. 2013;10(1):20-27. doi:10.3109/15412555.2012.727921.

31. Ware JH, Dockery DW, Louis TA, Xu XP, Ferris BG, Speizer FE. Longitudinal and cross-sectional estimates of pulmonary function decline in never-smoking adults. *Am J Epidemiol*. 1990;132(4):685-700. http://www.ncbi.nlm.nih.gov/pubmed/2403109. Accessed February 19, 2015.

32. Nishimura S, Betsuyaku T, Yoshioka H, Akiyama Y, Miyamoto K, Kawakami Y. Effect of aging on respiratory system. *Nihon Ronen Igakkai Zasshi*. 1996;33(3):825-828. doi:10.2147/ciia.2006.1.3.253.

33. Blanc PD, Torén K. Occupation in chronic obstructive pulmonary disease and chronic bronchitis: an update. *Int J Tuberc Lung Dis*. 2007;11(3):251-257. http://www.ncbi.nlm.nih.gov/pubmed/17352088. Accessed February 20, 2015.

34. Hnizdo E, Sullivan PA, Bang KM, Wagner G. Association between chronic obstructive pulmonary disease and employment by industry and occupation in the US population: a study of data from the Third National Health and Nutrition Examination Survey. *Am J Epidemiol*. 2002;156(8):738-746. http://www.ncbi.nlm.nih.gov/pubmed/12370162. Accessed February 20, 2015.

35. Abbey DE, Burchette RJ, Knutsen SF, McDonnell WF, Lebowitz MD, Enright PL. Long-term particulate and other air pollutants and lung function in nonsmokers. *Am J Respir Crit Care Med*. 1998;158(1):289-298. doi:10.1164/ajrccm.158.1.9710101.

36. De Marco R, Accordini S, Marcon A, et al. Risk factors for chronic obstructive pulmonary disease in a European cohort of young adults. *Am J Respir Crit Care Med*. 2011;183(7):891-897. doi:10.1164/rccm.201007-1125OC.

37. Silva GE, Sherrill DL, Guerra S, Barbee RA. Asthma as a risk factor for COPD in a longitudinal study. *Chest*. 2004;126(1):59-65. doi:10.1378/chest.126.1.59.

38. Vonk JM, Jongepier H, Panhuysen CIM, Schouten JP, Bleecker ER, Postma DS. Risk factors associated with the presence of irreversible airflow limitation and reduced transfer coefficient in patients with asthma after 26 years of follow up. *Thorax*. 2003;58(4):322-327. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1746641&tool=pmcentrez&rendertype=abstract. Accessed February 20, 2015.

39. Lange P, Parner J, Vestbo J, Schnohr P, Jensen G. A 15-year follow-up study of ventilatory function in adults with asthma. *N Engl J Med*. 1998;339(17):1194-1200. doi:10.1056/NEJM199810223391703.

40. Varraso R, Chiuve SE, Fung TT, et al. Alternate Healthy Eating Index 2010 and risk of chronic obstructive pulmonary disease among US women and men: prospective study. *BMJ*. 2015;350:h286. http://www.ncbi.nlm.nih.gov/pubmed/25649042. Accessed February 10, 2015.

41. Schols AM, Slangen J, Volovics L, Wouters EF. Weight loss is a reversible factor in the prognosis of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 1998;157(6 Pt 1):1791-1797. doi:10.1164/ajrccm.157.6.9705017.

42. US Department of Health and Human Services. The Health Consequences of Smoking-50 years of Progress: A report from the Surgeon Genera; 2014.

43. Prescott E, Lange P, Vestbo J. Socioeconomic status, lung function and admission to hospital for COPD: results from the Copenhagen City Heart Study. *Eur Respir J*. 1999;13(5):1109-1114. http://www.ncbi.nlm.nih.gov/pubmed/10414412. Accessed February 20, 2015.

44. Larson RK, Barman ML. The familial occurrence of chronic obstructive pulmonary disease. *Ann Intern Med*. 1965;63(6):1001-1008. http://www.ncbi.nlm.nih.gov/pubmed/5844558. Accessed March 5, 2015.

45. Lebowitz MD, Knudson RJ, Burrows B. Family aggregation of pulmonary function measurements. *Am Rev Respir Dis*. 1984;129(1):8-11. http://www.ncbi.nlm.nih.gov/pubmed/6703487. Accessed March 5, 2015.

46. Gottlieb DJ, Wilk JB, Harmon M, et al. Heritability of longitudinal change in lung function. The Framingham study. *Am J Respir Crit Care Med*. 2001;164(9):1655-1659. doi:10.1164/ajrccm.164.9.2010122.

47. Silverman EK, Palmer LJ, Mosley JD, et al. Genomewide linkage analysis of quantitative spirometric phenotypes in severe early-onset chronic obstructive pulmonary disease. *Am J Hum Genet*. 2002;70(5):1229-1239. doi:10.1086/340316.

48. Silverman EK. Genetic Epidemiology of COPD. *Chest*. 2002;121(3 Suppl):1S - 6S. http://www.ncbi.nlm.nih.gov/pubmed/11893649. Accessed February 26, 2015.

49. Hersh CP, Hokanson JE, Lynch D a., et al. Family history is a risk factor for COPD. *Chest*. 2011;140:343-350. doi:10.1378/chest.10-2761.

50. Wilk JB, Djousse L, Arnett DK, et al. Evidence for major genes influencing pulmonary function in the NHLBI Family Heart Study. *Genet Epidemiol*. 2000;19(July 1999):81-94. doi:10.1002/1098-2272(200007)19:1<81::AID-GEPI6>3.0.CO;2-8.

51. Silverman EK, Mosley JD, Palmer LJ, et al. Genome-wide linkage analysis of severe, early-onset chronic obstructive pulmonary disease: airflow obstruction and chronic bronchitis phenotypes. *Hum Mol Genet*. 2002;11(6):623-632. http://www.ncbi.nlm.nih.gov/pubmed/11912177. Accessed February 16, 2015.

52. Carrell RW. What we owe to alpha(1)-antitrypsin and to Carl-Bertil Laurell. *COPD*. 2004;1(1):71-84. doi:10.1081/COPD-120028703.

53. Silverman EK, Sandhaus RA. Clinical practice. Alpha1-antitrypsin deficiency. *N Engl J Med*. 2009;360(26):2749-2757. doi:10.1056/NEJMcp0900449.

54. DeMeo DL. 1:Antitrypsin deficiency 2: Genetic aspects of 1-antitrypsin deficiency: phenotypes and genetic modifiers of emphysema risk. *Thorax*. 2004;59(3):259-264. doi:10.1136/thx.2003.006502.

55. Pillai SG, Ge D, Zhu G, et al. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet*. 2009;5(3):e1000421. doi:10.1371/journal.pgen.1000421.

56. DeMeo DL, Mariani T, Bhattacharya S, et al. Integration of Genomic and Genetic Approaches Implicates IREB2 as a COPD Susceptibility Gene. *Am J Hum Genet*. 2009;85(4):493-502. doi:10.1016/j.ajhg.2009.09.004.

57. Wilk JB, Chen T-H, Gottlieb DJ, et al. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet*. 2009;5(3):e1000429. doi:10.1371/journal.pgen.1000429.

58. Cho MH, Boutaoui N, Klanderman BJ, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet*. 2010;42(3):200-202. doi:10.1038/ng.535.

59. Vestbo J, Anderson W, Coxson HO, et al. Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur Respir J*. 2008;31(4):869-873. doi:10.1183/09031936.00111707.

60. Fishman A, Martinez F, Naunheim K, et al. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med*. 2003;348(21):2059-2073. doi:10.1056/NEJMoa030287.

61. Zhu G, Warren L, Aponte J, et al. The SERPINE2 gene is associated with chronic obstructive pulmonary disease in two large populations. *Am J Respir Crit Care Med*. 2007;176(2):167-173. doi:10.1164/rccm.200611-1723OC.

62. Hancock DB, Eijgelsheim M, Wilk JB, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet*. 2010;42(1):45-52. doi:10.1038/ng.500.

63. Repapi E, Sayers I, Wain L V, et al. Genome-wide association study identifies five loci associated with lung function. *Nat Genet*. 2010;42(1):36-44. doi:10.1038/ng.501.

64. Soler Artigas M, Loth DW, Wain L V, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet*. 2011;43(11):1082-1090. doi:10.1038/ng.941.

65. Imboden M, Bouzigon E, Curjuric I, et al. Genome-wide association study of lung function decline in adults with and without asthma. *J Allergy Clin Immunol*. 2012;129(5):1218-1228. doi:10.1016/j.jaci.2012.01.074.

66. Hansel NN, Ruczinski I, Rafaels N, et al. Genome-wide study identifies two loci associated with lung function decline in mild to moderate COPD. *Hum Genet*. 2013;132(1):79-90. doi:10.1007/s00439-012-1219-6.

67. Tang W, Kowgier M, Loth DW, et al. Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function. *PLoS One*. 2014;9(7). doi:10.1371/journal.pone.0100776.

68. Lutz SM. A Genome-wide Associtaion Study Idenitifes Novel Risk Loci for Spriometric Measures among Smokers of European and African Ancestry. *unpublished*.

69. Parker MM, Foreman MG, Abel HJ, et al. Admixture Mapping Identifies a Quantitative Trait Locus Associated with FEV1 /FVC in the COPDGene Study. *Genet Epidemiol*. 2014:1-8. doi:10.1002/gepi.21847.

70. Hancock DB, Artigas MS, Gharib S a., et al. Genome-Wide Joint Meta-Analysis of SNP and SNP-by-Smoking Interaction Identifies Novel Loci for Pulmonary Function. *PLoS Genet*. 2012;8(12). doi:10.1371/journal.pgen.1003098.

71. Wain L V, Soler Artigas M, Tobin MD. What can genetics tell us about the cause of fixed airflow obstruction? *Clin Exp Allergy*. 2012;42(8):1176-1182. doi:10.1111/j.1365-2222.2012.03967.x.

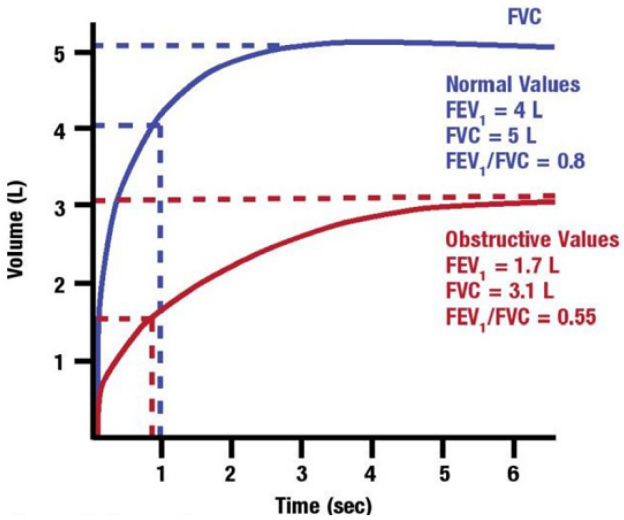


Figure 1.1
Volume to time curve for a normal versus obstructed airflow from spirometry. FEV_1 is the forced expiratory volume in one second. FVC is the forced vital capacity. A FEV_1/FVC ratio of less than 0.7 defines COPD (Figure adapted from the Global Initiative for COPD report[3]).

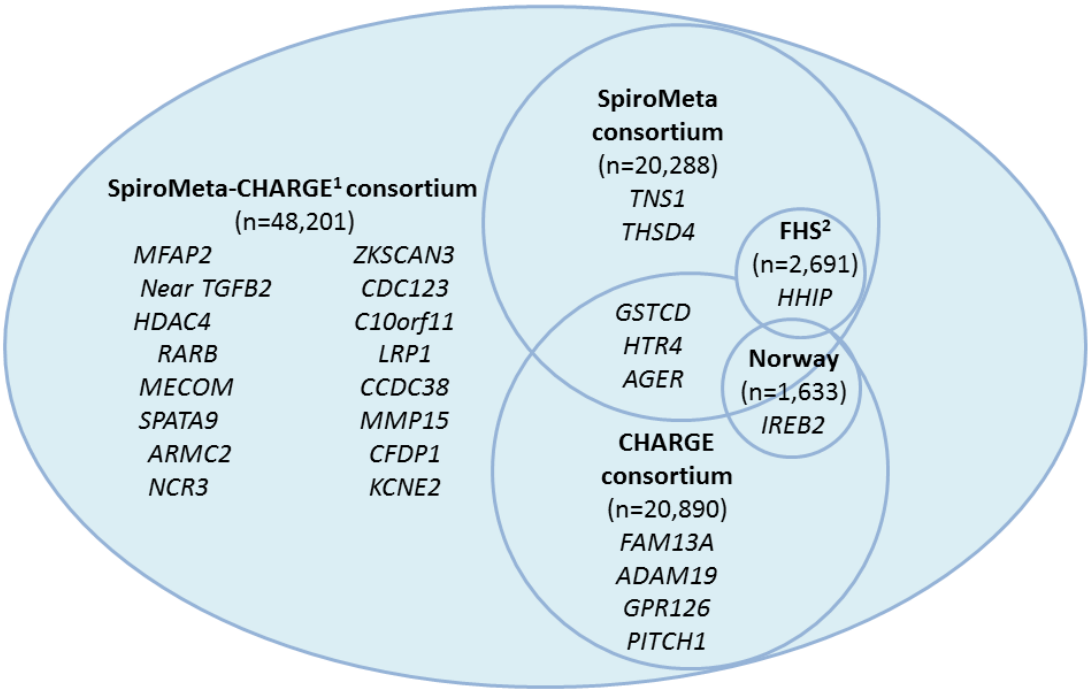


Figure 1.2
Genes associated with lung function (measured as FEV_1 or FEV_1/FVC) in European-derived populations (Figure adapted from Wain et al., 2012[71]).

Definition of abbreviations: [1]CHARGE= Cohorts for Heart and Aging Research in Genetic Epidemiology, [2]FHS = Framingham Heart Study

Table 1.1 Summary of published GWAS studies of lung function decline.

Author	Year	Study	Outcome(s) tested	Associated Genes(chr)
Imboden et al.	2012	EGEA[1], SAPALDIA[2], ECRHS[3]	change in FEV_1, change in FEV_1/FVC	DLEU7 (13)
Hansel et al.	2013	LHS[4]	change in FEV_1	TMEM26(10), FOXA1 (14), ANK3 (10)
Tang et al	2014	CHARGE[5] or Spriometa Consortium Studies	change in FEV_1	IL16/STARD5/TMC3 (15), ME3 (11)

Definition of abbreviations: [1]EGEA = Epidemiological Study of the Genetics and Environment of Asthma , [2]SAPALDIA = Swiss Cohort Study on Air Pollution and Lung and Heart Disease in Adults , [3]ECRHS = European Community Respiratory Health Survey, [4]LHS = Lung Health Study, [5]CHARGE = Cohorts for Heart and Aging Research in Genetic Epidemiology

Chapter 2. Exome Array Quality Control

Chapter 2. Exome Array Quality Control

Introduction

Genome-wide association studies (GWAS) have successfully identified common variation related to many complex phenotypes, but typically associated variants account for only a small fraction of the estimated heritability. Rare variants (defined here as variants with a minor allele frequency less that 5%) are hypothesized to play an important role in complex traits but are poorly characterized by GWAS arrays. Sequencing, either with whole exome or genome sequencing, can directly assay these rare variants but it is expensive and data interpretation is challenging given the large amount of data sequencing produces. Therefore, the Illumina HumanExome genotyping array (or "the exome array") was developed to capture known protein coding variation at a relatively low cost.

The exome array contains over 240,000 markers chosen from approximately 12,000 sequenced exomes based on the following criteria: 1) non-synonymous variants if observed at least three times in two or more studies; 2) stop-altering variants if observed at least two times in two or more studies; and 3) splice site variants if observed at least two times in two or more studies[1]. The array also contains 3,241 ancestry informative markers, 4,761 GWAS tag markers, 3,369 identity-by-descent markers, and 4,651 randomly selected synonymous markers to aid in quality control[1]. Previous research comparing the exome array to exome sequencing indicates that the coverage of rare, functional variants on the array is around 70%[2–5] and is comparable between European-derived and African-derived populations[4].

Although the exome array affords an exciting opportunity to assess rare coding variation, it also introduces novel data processing challenges. Unlike sequencing, genotyping platforms depend on automated clustering algorithms (e.g. Illumina's GenCall) to detect and assign genotype calls. These algorithms, originally designed for common variant detection, have difficulty accurately making genotype calls for rare variants because few observations exist in the heterozygote and homozygote minor allele clusters[6,7]. This can result in misclassified data (i.e. the incorrect assignment of a genotype to a cluster) or missing data (i.e. the inability to assign a genotype to a cluster). Additional sources of variability in exome array genotype calling include: 1) the sample size available for clustering (larger sample sizes improve genotype calling accuracy[5]); and 2) the manual re-clustering steps required by many exome array protocols, which can be subjective[6].

This chapter will describe the quality control (QC) procedures completed in exome array data from the COPDgene study, an observation case-control study of smokers with and without COPD[8]. A total of 9,858 Non-Hispanic White (NHW) and African American (AA) participants were genotyped on Illumina's HumanExome platform in addition to completing detailed questionnaires, pre- and post- bronchodilator spirometry, and volumetric computed tomography (CT) scans. This rich phenotyping, paired with the exome array data genotyping, provide a unique opportunity to extend the allelic and functional spectrum of genetic variation underlying COPD and COPD-related phenotypes. However, adding to the data processing challenges of the exome array, the 9,858 COPDgene subjects were genotyped in two batches, introducing the possibility of batch effects to contend with during the QC process (2,306 subjects on HumanExome array version 1.1 at University of Washington and 7,552 subjects on array version 1.2 at the Center for Inherited Disease Research). With this in mind, we undertook data

cleaning with the goal of producing a set of high quality functional variants for use by COPDgene investigators.

This chapter aims to: 1) provide a comprehensive description of the COPDgene exome array quality control procedures and; 2) to summarize variants available for analysis with regard to frequency and functional annotation.

Methods

Sample selection

All eligible participants of the COPDgene study with available DNA (n=9,858) were genotyped on Illumina's HumanExome Beadchip in two batches. Samples in batch one were selected based on extreme phenotypes and included NHW subjects with severe COPD and NHW disease-resistant smoking controls (n=2,306). Samples in batch two included the remaining NHW subjects (not genotyped in batch one) and all AA subjects (n=7,552). Therefore, in NHW subjects, chip membership is phenotypically determined (Table 1).

Exome array genotyping

Batch one was genotyped using version 1.1 of the HumanExome Beadchip (Illumina) at the University of Washington. Batch two was genotyped using version 1.2 of the HumanExome Beadchip (Illumina) at the Center for Inherited Disease Research at Johns Hopkins University. Each batch contained replicate samples (n= 128 in batch 1, n=38 in batch 2) and 121 samples were genotyped on both exome arrays to assess batch effects. In addition, batch two contained 166 control samples from 8 HAPMAP populations.

Because of differences between array version 1.1 and 1.2, we were unable to pool the two batches for joint genotyping calling. Thus, genotypes were called separately by

batch using Illumina's GenTrain clustering algorithm (version 1.0) in GenomeStudio (version 2011.1).

Exome array QC overview

An overview of the COPDgene exome array QC is provided in Table 2. Briefly, we performed: 1) initial single nucleotide variant (SNV) QC in order to combine the two exome array datasets 2) sample QC 3) SNV QC; and lastly 4) separated NHW and AA subjects for data analysis. All exome array QC was performed in using PLINK version 2.0[9] or R[10].

Initial SNV QC

Because marker identifiers on different exome array versions are inconsistent, we mapped all available SNVs to dbSNP version 141. In this process, we removed indels (n=276), markers with probable strand issues (n=135), markers that mapped to the wrong variant type (e.g. mapped to indel but was a SNV) (n=2), multi-allelic markers (n=2,680), and markers that did not map to dbSNP141 (n=617). Additionally, the Illumina exome array contains duplicate markers (n=833 on version 1.1, n=815 on version 1.2). For each duplicate marker, we checked concordance, dropped any discordant markers (n=107), and dropped the SNV with more missingness for concordant markers (n=1,541).Together, this initial SNV QC enabled us to merge the two exome array datasets for sample QC. Table 2.1 summarizes the initial SNV QC process.

Sample QC

A summary of sample QC is provided in Table 2.2.

GWAS concordance

All but 5 samples had been previously genotyped with Illumina's OmniExpress GWAS array and thus we could rely on previously performed QC for many of the sample checks

as long as concordance between samples was established. Therefore, we calculated genotype concordance between exome array samples and GWAS samples on SNVs that overlapped between the two platforms (n= >17,000 SNVs). This was performed separately for the first and second chips, as some samples were known to be GWAS discordant on the first chip and were re-genotyped on the second chip for that reason. Subjects with greater than 500 discordant sites between the two arrays were removed from analysis, excluding a total of 13 subjects.

Call rate

A low call rate, or fraction of called SNVs per sample, can indicate poor sample quality. We calculated call rate per sample using PLINK. This was performed by chip, as some markers were unique to either the first chip (n=3,219) or the second chip (n=406). Using a cutoff of 95%, no samples were dropped for low call rate.

Sex check

We relied on previously performed GWAS QC to exclude subjects with sex discrepancies. This excluded one subject. Additionally, there were 7 subjects with sex chromosome abnormalities identified in GWAS analysis that were genotyped on the exome chips (3 XO individuals, 3 XXY individuals, and 1 XY individual that self-identified as a female). These subjects were used in GWAS analysis and are therefore included in the exome chip dataset.

Heterozygosity

Given a homogenous sample, the heterozygosity rate, or fraction of non-missing genotype calls that are heterozygous, can help identify problematic samples (low heterozygosity may indicate inbreeding, while high heterozygosity may sample contamination). We calculated the heterozygosity using PLINK, separately in NHW and

AA subjects. Using a cutoff of greater or less than 6 standard deviations from the mean, 8 NHW and 4 AA outliers were identified and excluded from analysis.

Relatedness

To identify cryptic relatedness, we estimated kinship coefficients between all pairwise samples using the Kinship-based Inference for GWAS (KING) software[11] (using the option: "-kinship -related -degree 2"). All samples were estimated together (excluding known duplicates), using all available SNVs. A total of 152 second degree relationships were identified by KING, of which 122 had been previously identified and were removed (114 removed during GWAS QC process and 8 removed during GWAS concordance checks). Additionally, 23 relationships (representing 7 subjects), were removed for suspected contamination. Seven additional relationships identified as possible second degree relationships by KING (kinship coefficient 0.88-0.95), but not flagged during the GWAS QC process, were retained.

Population outliers

To identify population outliers, we ran principal components analysis (PCA) separately in NHWs and AAs using unlinked, autosomal polymorphisms (n=14,961 SNPs in NHW and 19,884 SNPs in AAs). These results (using markers from the exome chip) were nearly identical to the CPA results using markers from the GWAS chip, and therefore we excluded subjects identified as population outliers in the GWAS analysis (n=52) (Figure 2.3).

Replicate and control samples

Replicate samples were genotyped on the first chip (n=128), the second chip (n=38), and across the 2 chips (n=121). For all replicates, we calculated genotype concordance

between the samples and dropped the one with more missingness. Additionally, we excluded 166 HAPMAP genotyping control samples included on chip # 2.

Fraction of singleton/doubletons

An excess or depletion of very rare variants can indicate sample quality issues including sample contamination, inbreeding or population outliers. Thus, we looked at the number of singletons and doubletons per person (separate in NHW and AAs), but did not exclude any subjects based on this metric (Figure 2.4).

Additional GWAS exclusions

Seven subjects were excluded from analysis for having confirmed alpha-1 antitrypsin deficiency. One subject was flagged as mislabeled during the GWAS QC process, and therefore was also excluded from the exome array analysis.

Single Nucleotide Variant (SNV) QC

A summary of SNV QC is provided in Table 2.3.

Call rate

We assessed SNV call rate using PLINK. Because there are some markers unique to only one chip, this was performed separately by chip. Using a 95% call rate cutoff, a total of 1,394 SNVs were dropped from the first chip, and a total of 4,874 SNVs were dropped from the second chip.

Frequency differences between chips

To test genotype quality, we compared the allele frequencies of overlapping markers on chip v1.1 to allele frequencies on chip v1.2 in NHW subjects. We used two frequency mismatch cutoffs to exclude markers: 1) SNVs with absolute frequency differences greater than 0.1; and 2) SNVs that were significantly different on a Fisher's exact test of variant frequency between controls on each chip (cutoff = p-value < 10^{-4}). In total, this

excluded 180 SNVs. Additionally, we flagged (but did not exclude) markers with p-values between 10^{-3} and 10^{-4} on the Fisher's exact test of frequency differences.

Hardy-Weinberg equilibrium

We calculated Hardy-Weinberg equilibrium (HWE) by race in unaffected (FEV_1/FVC > 0.70) subjects using PLINK. Markers with HWE p-values less than 10^{-8} were excluded (n=150 in NHW, n=143 in AAs). Markers with HWE p-values less than 10^{-4} were flagged but not excluded from the final dataset. Manhattan plots of HWE p-values by race are provided in Figure 2.5.

Minor allele concordance

Using the duplicate samples within each chip (n=166) and between the two chips (n=121), we calculated minor allele concordance (MAC) and dropped SNVs with MAC < 95% (n=1,677).

Non-autosomal markers

A total of 5,540 markers on the X, Y, XY, and mitochondrial chromosomes were removed from analysis. These SNVs were not included in the QC process and require additional QC if used in future analyses.

Chip-specific markers

There are markers unique to only chip one (n=3,219) or chip two (n=406). Markers genotyped on only 1 of the 2 chips are included in the analyzed dataset, but we created a flag for consideration during analysis as these SNVs will have will substantial missingness which may be non-random (chip membership is phenotype-dependent in NHW subjects).

Annotation

SNVs were annotated using ANNOVAR[11] with the RefSeq reference genome[12].

Results

A total of 6,581 subjects and 233,263 SNVs passed QC in NHWs, and 3,221 subjects and 233,255 SNVs passed QC in AAs. The overall call rate on chip one was 99.8% (sample range: 98.1% - 100.0%). The overall call rate on chip two was 98.0% (sample range: 95.4% – 98.0%).

Minor allele frequencies and functional annotation

The minor allele frequency (MAF) distributions for successfully genotyped SNVs are described (by race) in Table 2.4. There were 82,832 monomorphic SNVs in NHWs (35.5%) and 86,605 monomorphic SNVs in AAs (37.1%). The majority (> 87%) of variants in both NHWs and AAs have a MAF < 5% and many (> 37%) have a MAF < 0.005%. NHW subjects have more variants with a MAF less than 0.001 (90,946 in NHW vs. 55,132 in AAs), likely attributable to the marker contents of the chip (chosen from primarily NHW exomes) and to the number of AAs genotypes (about half the number of NHWs).

Greater than 92% of genotyped SNVs in both NHWs and AAs are functional (annotated as nonsynonomous, stopgain, stoploss, or splicing). The most common function category is nonsynonomous, as it accounts for 90% of all genotyped SNVs. A summary of the functional categories of observed SNVs by MAF is provided in Table 2.5.

Genes with multiple SNVs

Given that the vast majority of observed variants are rare, traditional approaches testing for the association of a single variant and outcome are underpowered. To address this, many have suggested different collapsing methods where sets of rare variants in a given genomic unit (e.g. gene) are “collapsed” and collectively tested for an association with the outcome. This test is more powerful than a single variant test if the collapsed

variants are causal. To assess our ability to perform “collapsing” tests by gene, we summarized the number of variants in each gene by MAF category (Table 2.6).

There are 16,142 genes in NHWs and 15,932 genes in AAs with more than one observed variant. The median number of rare (MAF < 5%) variants per gene was 5 in both NHW and AAs. Three genes had an usually high number (>100) of observed variants *TTN*, *MUC16* and *OBSCN*, but these genes are large and known to contain a high number of variants. Together, these analyses indicate that the majority of genes observed in the COPDgene exome array data have multiple rare variants, and are therefore eligible to be used in “collapsing” tests.

Coverage of COPD-associated genes

Previous research has been successful in identifying common variation associated with COPD through genome-wide association studies (GWAS)[13]. While GWAS-associated SNPs may identify important genetic regions, they are likely not causal and only tag functional mutations. Because the exome array directly queries functional variants, we have the opportunity to potentially elucidate causal mutations explaining previously identified GWAS signals. Thus, we assessed the number of observed variants by MAF category in 26 genes associated with lung function in a recent large meta-analysis of 48,201 NHW subjects[14] (Figure 2.6).

The number of variants per lung function-associated gene ranged from (0-50). One gene (*CDC123*) was not observed in our data. As expected, the majority of SNVs in lung function-associated genes were rare and functional.

Batch Effects

The COPDgene NHW subjects were genotyped at two time points, with different versions of the exome array. However, 121 subjects were genotyped on both chips. We used these duplicated samples to assess batch effects.

Figure 2.7 shows variant counts in different frequency categories (e.g. singleton, doubleton) in the 121 duplicates samples. If no batch effect were present, we would expect counts in each category to be equal. However, we see that there are slightly fewer variants called and slightly more monomorphic markers in the first chip than the second chip, suggesting there may be some under calling of rare variants in chip one. Potential reasons for this include: 1) differences in genotype clustering and calling between the 2 arrays; and 2) the larger sample size used for calling chip two.

We also calculated minor allele concordance (the number of concordant minor alleles over the total number of minor alleles) between duplicated samples. Overall the minor allele concordance between samples genotyped on both chips was high (> 99%) for all frequencies (Table 2.7). Together, this suggests there are subtle but present batch effect in the NHW exome array data.

Discussion

This chapter describes the quality control procedures completed in the COPDgene exome array data. Through multiple subject and variant QC steps, we produced a high quality dataset to be tested for associations with lung function and the wide array of COPD-related outcomes available in the COPDgene study. We have shown that this data consists of mostly rare and potentially protein altering variants, including many genes with multiple rare variants and genes previously associated with lung function. This data provides researchers with an opportunity to extend the allelic and functional

spectrum of genetic variation underlying COPD, and this chapter serves as a resource for those using it.

References

1. Exome Chip Design Wiki. 2013. genome.sph.umich.edu/wiki/Exome_Chip_Design. Accessed on 7/5/2013.

2. Peloso GM, Auer PL, Bis JC, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet*. 2014;94(2):223-232. doi:10.1016/j.ajhg.2014.01.009.

3. Holmen OL, Zhang H, Fan Y, et al. Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. *Nat Genet*. 2014;46(4):345-351. doi:10.1038/ng.2926.

4. Igartua C, Myers RA, Mathias RA, et al. Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. *Nat Commun*. 2015;6:5965. doi:10.1038/ncomms6965.

5. Grove ML, Yu B, Cochran BJ, et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. *PLoS One*. 2013;8(7):e68095. doi:10.1371/journal.pone.0068095.

6. Guo Y, He J, Zhao S, et al. Illumina human exome genotyping array clustering and quality control. 2014;9(11):2643-2662. doi:10.1038/nprot.2014.174.

7. Wang GT, Peng B, Leal SM. Variant association tools for quality control and analysis of large-scale sequence and genotyping array data. *Am J Hum Genet*. 2014;94(5):770-783. doi:10.1016/j.ajhg.2014.04.004.

8. Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDgene) study design. *Epidemiology*. 2011;7(1):1-10. doi:10.3109/15412550903499522.Genetic.

9. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1):7. doi:10.1186/s13742-015-0047-8.

10. R Core Team. R: A language and environment for statistical computing. *R Found Stat Comput*. 2013.

11. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164. doi:10.1093/nar/gkq603.

12. Pruitt KD, Brown GR, Hiatt SM, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*. 2014;42(Database issue):D756-D763. doi:10.1093/nar/gkt1114.

13. Soler Artigas M, Loth DW, Wain L V, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet*. 2011;43(11):1082-1090. doi:10.1038/ng.941.

14. Soler Artigas M, Wain L V., Tobin MD. Genome-wide association studies in lung disease. *Thorax*. 2012;67:271-273. doi:10.1136/thoraxjnl-2011-200724.

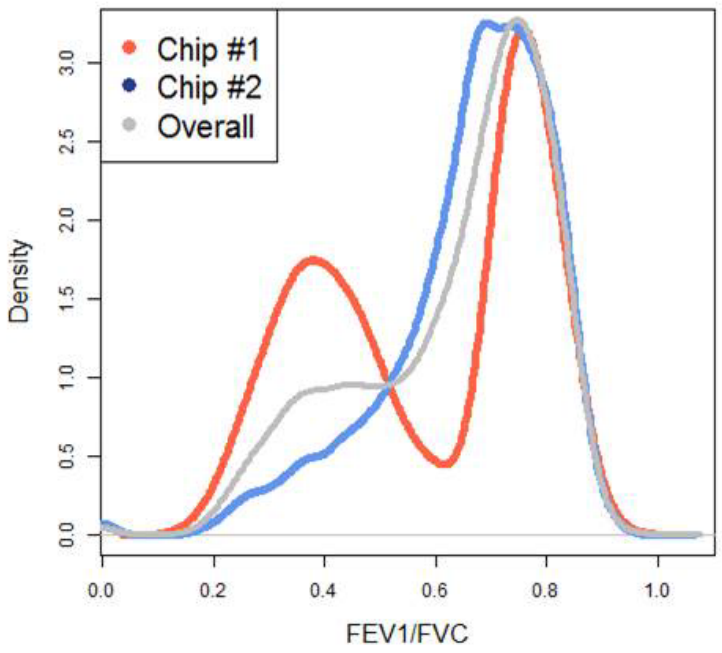


Figure 2.1
Distribution of FEV_1/FVC by chip in NHW COPDgene subjects.Chip one samples were selected based on extreme phenotypes (severe COPD cases and disease-resistant smoking controls). Chip two samples included all other COPDgene NHW subjects.

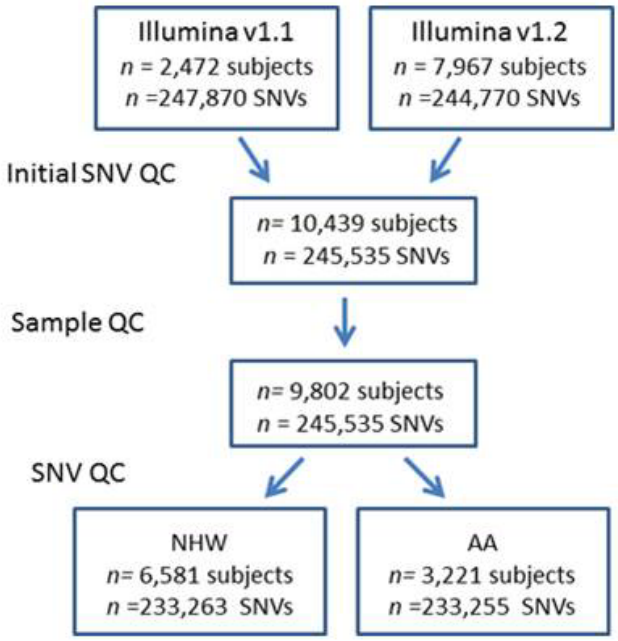


Figure 2.2
Overview of COPDgene exome array quality control process, including the number of subjects and SNVs used for each QC step. We performed: 1) initial SNV QC; 2) Sample QC 3) SNV QC; and 4) separated NHW and AA subjects for data distribution.

Table 2.1
Summary of initial SNV quality control. We mapped all SNVs to dbSNP 141 to create consistent marker names. During this process, we dropped indels, SNVs with probable strand issues, markers that were the wrong type (SNV versus indel), multi-allelic markers, duplicate markers, and SNVs that were unable to be mapped to dbSNP141.

Flag	N SNVs dropped Chip #1	N SNVs dropped Chip #2
Indel	140	136
Strand Problem	50	85
Wrong marker type	1	1
Multi-allelic	1340	1340
Duplicate Marker	865	890
Not mapped to dbSNP	473	144

Table 2.2
Summary of subject QC in the COPDgene exome array data. A total of 662 subjects were flagged (including 27 subjects flagged more than once) resulting in 9,802 subjects available for analysis.

Flag	Cutoff	Number of Subjects
GWAS discordant	> 500 discordant sites	13
Call Rate by Subject	< 95%	0
Sex Check	GWAS exclusions	1
Heterozygosity	> |6 SD| from mean	12
Relatedness	GWAS exclusions and samples related to ≥ 3 people	111
Population outliers	GWAS exclusions	42
GWAS excluded	Alpha 1 Anti-trypsin or mislabeled	8
Known Duplicates	Sample with less missingness	287
Known Duplicate Sample flagged	Failed one of above flags	22
HAPMAP controls		166

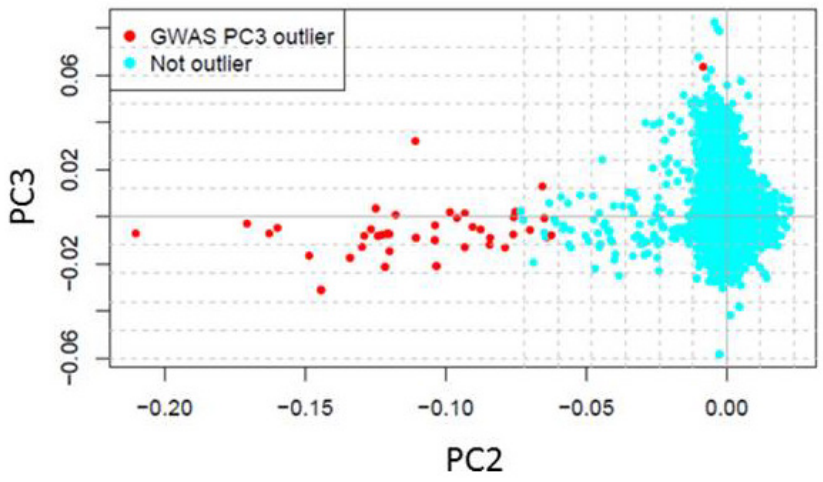


Figure 2.3
Comparison of principal components generated from NHW exome array data and NHW GWAS data. The x-axis is principal component 2. The y-axis is principal component 3. A total of 52 subjects were excluded as population outliers (colored red).

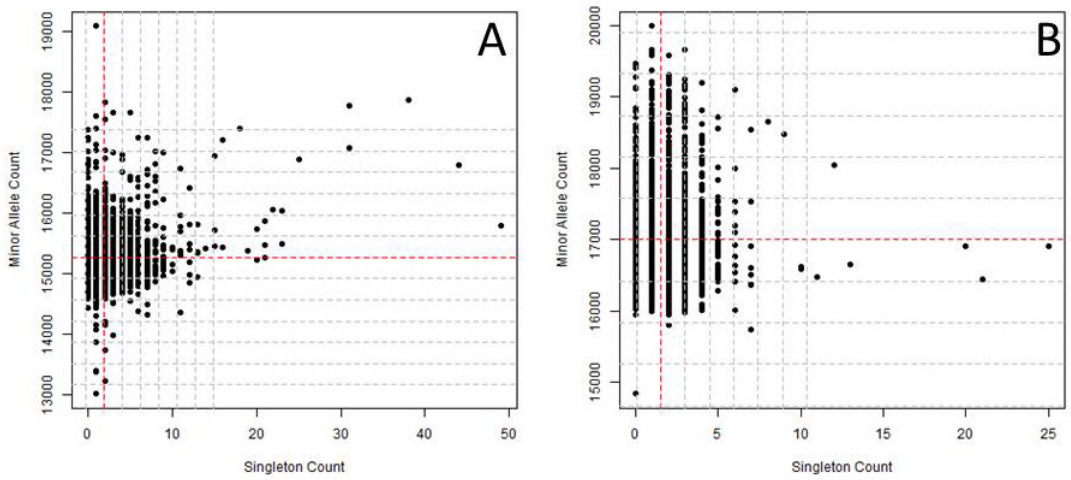


Figure 2.4
Singleton count versus total minor allele count in A) NHW COPDgene subjects (n=6,581) and B) AA COPDgene subjects (n=3,321). Red lines indicate mean counts, gray line indicate standard deviations. As expected, NHW subjects have fewer total minor alleles than AA subjects (mean NHW = 15,270, mean AAs = 17,000). Outliers in counts of singletons tend to match up with principal component outliers.

Table 2.3
Summary of SNV QC in the COPDgene exome array data.

Flag	Cutoff	Number of SNVs dropped
Call rate	< 95% by chip	Chip #1: 1,394 Chip #2: 4,874
Frequency differences between chips	MAF difference > 0.1 Fisher exact test P < 10^{-4}	180
Hardy- Weinberg Equilibrium	Race specific P < 10^{-8} in control samples	NHW: 150 AA: 143
Minor allele concordance	< 95% in replicate samples	1,677
Non-autosomal markers	X, Y, XY, MT	5,540

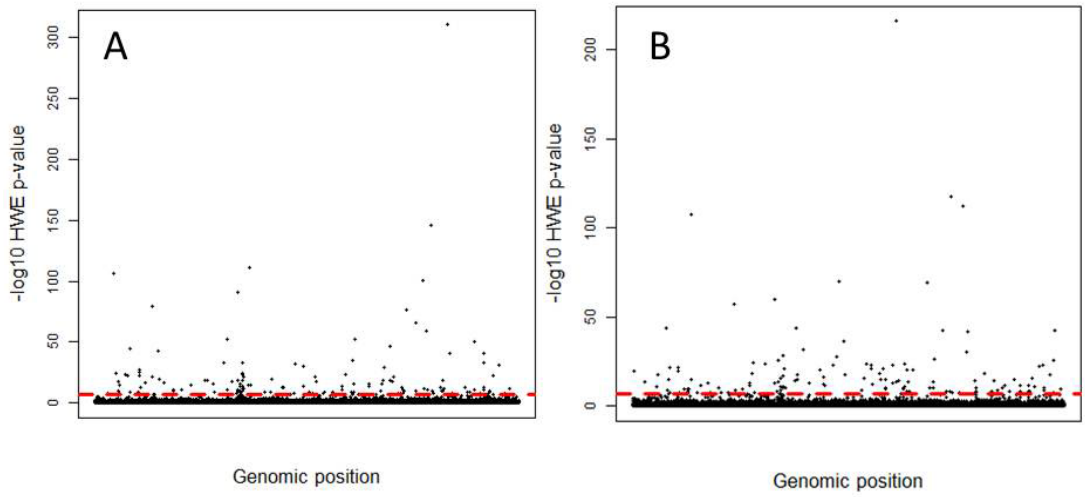


Figure 2.5
Manhattan plots of Hardy Weinberg p-values for A) NHW controls ($FEV_1/FVC > 0.70$) and B) AA controls ($FEV_1/FVC > 0.70$). The dashed red line denotes the p-value cutoff for exclusion ($p < 10^{-8}$).

Table 2.4
Minor allele frequency distributions by race.

MAF Interval	Non-Hispanic Whites (n=6,581) (%)	African Americans (n=3,221) (%)
0	82,832 (35.5)	86,605 (37.1)
(0, 0.001]	90,946 (39.0)	55,132 (23.6)
(0.001, 0.005]	19,654 (8.4)	32,933 (14.1)
(0.005, 0.01]	5,305 (2.3)	10,224 (4.4)
(0.01, 0.05]	9,167 (3.9)	18,715 (8.0)
(0.05, 0.1]	4,143 (1.8)	6,426 (2.8)
(0.1, 0.2]	5,800 (2.5)	7,035 (3.0)
(0.2, 0.5]	15,630 (6.8)	16,351 (7.0)

Table 2.5
Counts of observed variants by functional type in COPDgene NHW and AAs. SNVs were annotated using ANNOVAR with RefSeq reference genome.

Non-Hispanic White (n= 6,581)

Variant Type	All Sites	MAF < 1 %	MAF 1-5%	MAF > 5%
Nonsynonymous	122,656	106,344	7,554	8,758
Stopgain/Stoploss	2,538	2,395	73	70
Splicing	1,074	983	36	55
Synonymous	8,521	4,507	562	3,452
Other	1,318	982	98	238

African American (n= 3,221)

Variant Type	All Sites	MAF < 1 %	MAF 1-5%	MAF > 5%
Nonsynonymous	119,407	90,893	16,463	12,051
Stopgain/Stoploss	2,003	1,760	150	93
Splicing	873	718	75	80
Synonymous	8,761	3,735	1,126	3,900
Other	1,320	846	191	283

Table 2.6
Counts of observed variants by gene. The majority of genes have more than one variant and are therefore eligible to be included in a collapsing test.

	Non-Hispanic Whites	African Americans
N genes with observed variant(s)	18,053	17,924
N genes with 1 observed variant	1,911	1,992
Median N variants per gene (range)	6 (1-572)	6 (1-558)
Median N variants MAF < 5% per gene (range)	5 (1-541)	5 (1-496)
Median N variants MAF < 1% per gene (range)	5 (1-483)	4 (1-434)

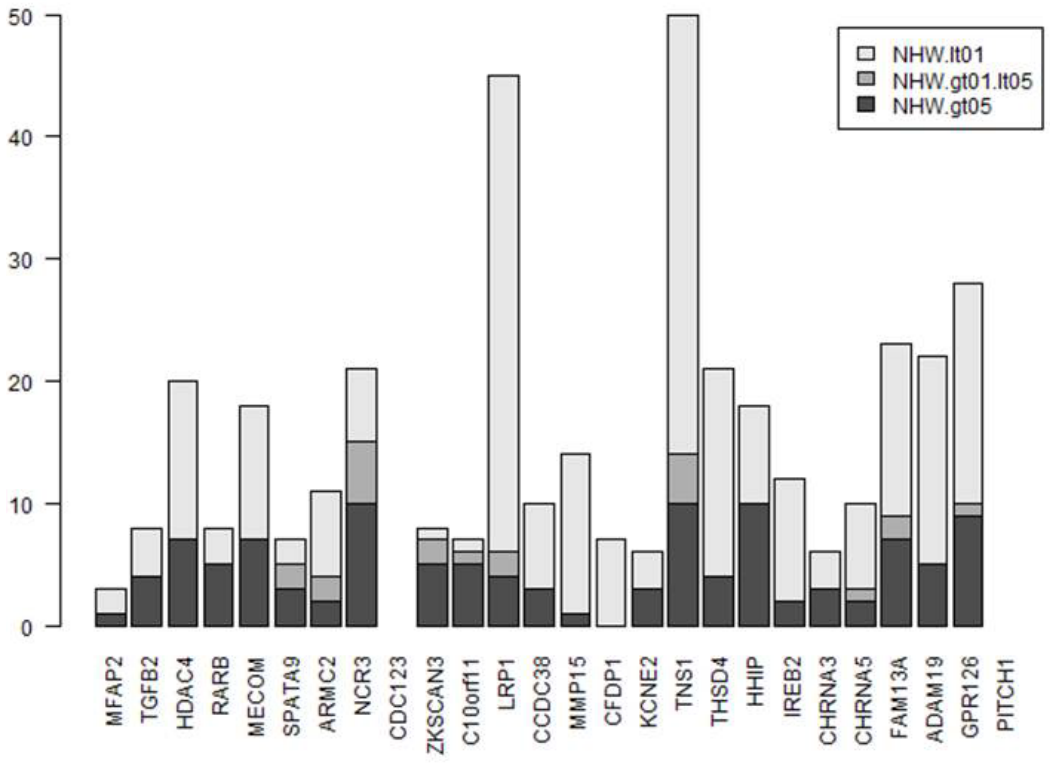


Figure 2.6
Counts of observed variants in known COPD-associated genes in NHW COPDgene exome array. COPD-related genes (n=26) from a large meta-analysis of 48,201 NHW subjects[14]. Dark gray bars indicate counts of variants with MAF > 5%, gray bars indicate counts of variants with a MAF between 1% and 5%, and light gray bars counts of variants with MAF < 1%.

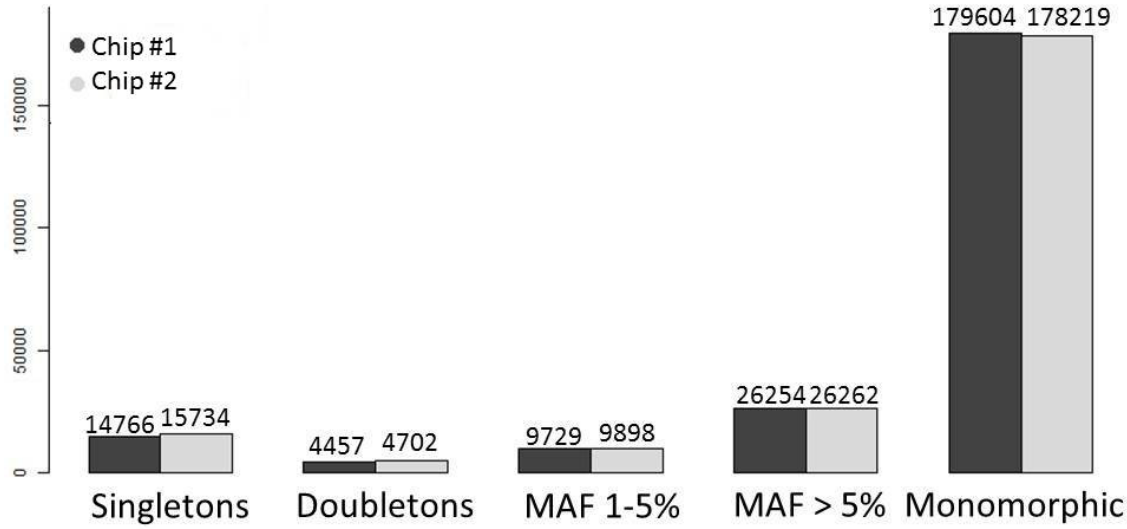


Figure 2.7
Counts of variants by frequency category in 121 duplicated samples. Singletons are variants observed once in the dataset, doubletons are variants observed twice in the dataset. Overall, there are fewer variants called on the first chip than the second chip in these duplicated samples.

Table 2.7
Minor allele concordance in 121 subjects genotyped on both exome array #1 and exome array #2. Overall minor allele concordance is high (>99%) in all frequency categories.

	N markers	Concordance %
Singletons	15,734	99.39
Doubletons	4,702	99.36
MAF 1- 5%	9,898	99.79
MAF > 5%	26,262	99.86
Overall	56,596	99.64

Chapter 3. Exome Array Analysis of Lung Function in the COPDgene Study

Chapter 3. Exome Array Analysis of Lung Function in the COPDgene Study

Introduction

Chronic obstructive pulmonary disease (COPD) is a progressive respiratory disease characterized by airflow obstruction and decreased lung function. It is the 3rd leading cause of death worldwide, accounting for approximately 3 million deaths in 2010[1]. COPD is diagnosed based on spirometric measures of lung volume and flow, specifically the ratio of forced expiratory volume in one second (FEV_1) to forced vital capacity (FVC). The FEV_1/FVC ratio reflects the severity of airway obstruction and can predict morbidity and mortality in individuals[2].

The primary risk factor for COPD is cigarette smoking, but genetics also play a role in individual susceptibility and disease progression[3]. Over the past several years, substantial advances have been made in our understanding of the genetics underlying COPD using genome wide association studies (GWAS) and large-scale meta analyses[4–8]. However, associated genetic risk factors account for only a small fraction of the estimated heritability of lung function[8], suggesting much of the genetic variation in spirometric measures has yet to be discovered.

Some of this "missing heritability" may reflect the effects of rare, protein coding variation, which cannot be easily tested using the GWAS approach. Nearly every gene contains functionally important rare variants[9] and recently developed exome array technology, which directly types these variants, has been successful in identifying additional loci associated with many complex traits[10–14].

To determine if rare coding variation plays a role in lung function, we genotyped 9,858 participants from the COPDgene study[15] with Illumina's HumanExome Beadchip. We

performed both: 1) traditional single variant analysis, which tests for statistical association between one variant and an outcome; and 2) gene-based analysis, whereby all variants in a gene are "collapsed" and tested together for association with an outcome. These gene-based tests are more powerful than single variant tests whenever multiple variants in a gene are causal, and can be particularly useful for rare variants (minor allele frequency < 5%), where single variant tests offer little statistical power.

Methods

Study participants

We genotyped 9,858 participants of the COPDgene study, an observational study conceived in 2008 to investigate the genetic and environmental etiology of COPD. A complete study protocol for COPDgene had been described elsewhere[15]. Briefly, self-identified Non-Hispanic Whites (NHWs) and African Americans (AAs) between the ages of 45 and 80 years with a minimum of 10 pack-years smoking history were enrolled at 21 study centers across the United States. In addition to completing detailed questionnaires, pre- and post- bronchodilator spirometry, and volumetric computed tomography (CT) of the chest, participants provided whole blood for DNA extraction and genotyping.

Genotyping and quality control (QC)

Genotyping was performed with the Illumina HumanExome BeadChip v1.1 at the University of Washington or with v1.2 at the Center for Inherited Disease Research (CIDR). Genotype calling was performed using GenTrain v1.0 in GenomeStudio v2011.1 (Illumina). The quality control (QC) process resulted in several samples being dropped from the analysis. Sample exclusions included sex discordances (n=1), sample duplications (n=309), discordance with previous GWAS typing (n=13), population outliers

(n=42), unexpected related subjects (n=111) and samples with high overall heterozygosity (n=12). This resulted in a final sample size of 6,581 NHW and 3,221 AA subjects.

A total of 247,870 single nucleotide variants (SNVs) were genotyped on Illumina HumanExome BeadChip array v1.1, and a total of 244,770 SNVs were genotyped on v1.2. SNVs exclusions were based on call rate < 95% (n=6,268), Hardy-Weinberg $p < 10^{-8}$ (n=293), minor allele concordance between duplicate samples < 95% (n=1,677), non-autosomal SNVs (n=5,540) and significant frequency differences between the two genotyped arrays (n=180). This resulted in 150,299 SNVs in NHWs and 146,654 SNVs in AAs available for analysis. A complete description of all quality control procedures is provided in chapter 2 of this dissertation.

Phenotypes

We analyzed: 1) the ratio of forced expiratory volume in one second to forced vital capacity (FEV_1/FVC), and 2) the ratio of forced expiratory volume in one second as a percent of the predicted value (FEV_1 % predicted). After taking a maximal deep breath, FEV_1 is the volume of air exhaled in one second, while FVC is the total volume of air exhaled. The ratio of these two measurements (FEV_1/FVC) defines COPD when values are < 0.7. Forced expiratory volume in one second as a percent of the predicted value (FEV_1 percent predicted) is calculated using regression models incorporating race, gender, age and height, and is indicator of the severity of airflow obstruction[16]. Spirometry measurements were obtained using a standardized spirometer (EasyOne by ndd Medical Technologies) after administering two puffs (180 mcg) of albuterol.

Statistical analyses

Single variant analysis

We tested SNP-trait associations for FEV_1/FVC and FEV_1 percent predicted separately in NHWs and AAs assuming an additive genetic model using linear regression. All analyses were adjusted for age, gender, pack-years smoked, exome array version, and principal components (PCs) to control for population stratification within racial groups. PCs were calculated using unlinked, autosomal polymorphisms (n= 14,975 in NHW and n=19,498 AAs) in Eigensoft[17]. Single variant tests included only SNVs with a MAF > 0.5% annotated as being functional (nonsynonomous, stop-gain, stop-loss, splicing or ncRNA). Statistical significance for each quantitative phenotype was determined by Bonferroni correction ($0.05 \div 21{,}394$ tests $= 2.34 \times 10^{-6}$ in NHWs and $0.05 \div 38{,}519$ tests $= 1.30 \times 10^{-6}$ in AAs).

Gene-based analysis

We performed gene-based tests using the sequence kernel association test (SKAT)[18]. SKAT has been shown to perform well under various scenarios, including when protective, deleterious and null variants are present within a single gene. We tested putatively functional, rare (minor allele frequency [MAF] < 5%) variants in genes with ≥ 5 observed variants. Statistical significance was determined by Bonferroni correction for the number of genes tested in each population ($\alpha = 5.83 \times 10^{-6}$ for NHW analysis, $\alpha = 5.49 \times 10^{-6}$ for AA analysis).

For statistically significant genes, we tested the relative contribution of each variant by removing one SNV at a time, and re-calculating the evidence for association across the gene using the SKAT statistic. All analyses were adjusted for age, gender, pack-years smoked, exome array version and principal components to control for population stratification.

Variant annotation and *in-silico* functional prediction

SNVs were annotated using ANNOVAR[19] with the RefSeq[20] reference genome. In an effort to assess the potential functional consequences of variants identified as significant, we used the combined annotation dependent depletion (CADD) algorithm [21] to predict deleterious effect on the gene product and genomic evolutionary rate profiling (GERP) scores [22] to indicate the degree of evolutionary conservation of variants.

Results

Study participants

Clinical characteristics of the NHW (n=6,562) and AA (n=3,182) subjects with available spirometry data are provided in Table 1 and Figures S3.1-S3.2. On average, AA subjects were younger with fewer pack-years exposure and had higher mean FEV_1/FVC and FEV_1 percent predicted.

Genotypes

A total of 150,299 SNVs in NHWs and 146,654 SNVs in AAs passed QC and were not monomorphic. Greater than 87% of these were rare (MAF < 5%) and greater than 92% were putatively functional (annotated as nonsynonomous, stop-gain, stop-loss, ncRNA or splicing). The most common functional category was nonsynonomous, representing over 90% of all genotyped SNVs.

Single variant analysis

We identified four significant single variant associations with spirometry, all of which were nonsynonymous mutations (Table 3.2 and Figure 3.1). Three of these four associations had been previously identified in GWAS as genetic risk loci for lung function: rs16969968 in *CHRNA5* (Asp398Asn, MAF = 36.6%) and rs2070600 in *AGER* (Gly38Ser, MAF = 4.2%). The identified nonsynonomous variant in *CHRNA5* had a

phred-scaled CADD score of 10.5, indicating it was in the top 10% of deleterious predictions[21], and a GERP score of 3.88, indicating a high degree of evolutionary constraint (a cutoff of 2.0 is commonly used to indicate "constraint" [22]). Similarly, *in-silico* functional predictions using CADD and GERP suggest the identified nonsynonomous variant in *AGER* is likely deleterious (CADD score = 18.0, GERP score of 5.82).

We identified one novel association between a nonsynonomous SNV and FEV_1 percent predicted in AA COPDgene subjects (rs34664882, $p = 2.42 \times 10^{-7}$, MAF= 1.7%). The associated SNV is in the gene encoding the ankyrin 1 (*ANK1*) protein and the minor allele (A) is associated with significantly lower FEV_1 percent predicted ($\beta = -11.59$, Figure 3.2). This mutation causes the substitution of a valine for alanine in amino acid position 1503, however CADD and GERP did not predict this change to be deleterious (CADD score = 2.1, GERP score = -2.53). This variant (rs34664882) was observed in NHW subjects at a comparable frequency (MAF = 2.9%), but no association between genotype and FEV_1 percent predicted was observed in this sub-population (p=0.93). Manhattan and QQ plots for all single variant association tests are shown in Figures S3.3-S3.5.

Gene-based analysis

We identified three significant gene associations with FEV_1/FVC or FEV_1 percent predicted using SKAT (Table 3.3 and Figure S3.6). Associated genes include: *AGER* on chromosome 6 (associated with FEV_1/FVC in NHWs, $p = 2.01 \times 10^{-9}$); *ProSAPiP1* on chromosome 20 (associated with FEV_1 percent predicted in NHWs, $p = 2.63 \times 10^{-6}$); and *ANK1* on chromosome 8 (associated with FEV_1 percent predicted in AAs, $p = 5.27 \times 10^{-6}$). All gene-based associations were population-specific, and we observed no evidence of signal overlap between NHWs and AAs (SKAT p-value for *AGER* in AAs = 0.52, SKAT p-value for *ProSAPiP1* in AAs = 0.19, SKAT p-value for *ANK1* in NHWs = 0.93).

To quantify the relative contribution of each SNV to the gene-based signals, we removed one SNV at a time and re-calculated the evidence for association across the gene using SKAT (Figure 3.3). In each of the three associated genes, this identified a single, rare SNV that accounted for the gene-based signal (i.e. every time that variant was included in the re-calculated SKAT, gene-based results were significant, but when it was excluded, SKAT results were not significant). The identified SNV in *AGER* (rs2070600) and *ANK1* (rs34664882) was the same SNV previously identified in the single variant analysis. The identified SNV in *ProSAPiP1*, rs140282982, was observed in 10 subjects (MAF = 0.00076), and a single copy of the minor allele (C) reduced FEV_1 percent predicted by an average of -31.7 (Figure S3.7).

Discussion

We performed single variant and gene-based exome array analysis in NHW and AA participants from the COPDgene study to identify functional variation associated with quantitative spirometric phenotypes. We replicated previously known GWAS associations (rs2070600 in *AGER* and rs16969968 in *CHRNA5*), and we identified two novel associations in the *ANK1* and *ProSAPiP1* genes. Association signals in these genes were largely driven by a single rare, nonsynonomous variant with a large effect (rs34664882 in *ANK1* and rs140282982 in *ProSAPiP1*).

ANK1 is part of the ankyrin family of proteins that act to attach integral membrane proteins to the spectin-actin based cytoskeleton. It plays a key role in cell motility, cell activation, cell proliferation, and the maintenance of specific cellular membrane domains[20]. *ANK1* is highly expressed in the brain and muscle, and moderately expressed in the lung[23]. Mutations in *ANK1* are responsible for hereditary spherocytosis[24], and this gene (and others in the ankyrin family of proteins) have been associated with lung function in previous research: In 2012, Imboden et al. identified a

variant in *ANK1* (rs7006290) as associated with FEV_1 decline in asthmatics ($p=5.19 \times 10^{-6}$) [25]. Additionally, Hansel et al. identified an association in the *ANK3* gene with FEV_1 decline in 4,048 Lung Health Study participants [26].

ProSAPiP1, or the proline-rich synapse-associated protein-interacting protein 1 (also known as *LZTS3)*, is thought to play a role in regulating cellular growth, although its biological function remains largely unknown[27]. This gene is primarily expressed in the brain and kidney, but shows moderate expression in multiple other tissues including the lung[23]. There are no known associations between this gene and clinically defined COPD or COPD-related phenotypes, and its role in lung disease remains unclear.

Our results indicate low frequency, coding variants may account for some known GWAS signals, and association signals in newly identified genes may explain a small proportion of the population variation in lung function. However, overall we found no widespread impact of low-frequency coding variation in lung function. One limitation of our study is that the exome array does not cover all functional variants. Specifically, very rare variants (such as uncharacterized private mutations) and regulatory variation outside of the exome are not genotyped on this platform, and whole genome sequencing may be needed for a comprehensive assessment of al variation. We have identified interesting results warranting further research to understand the mechanism underlying their role in COPD etiology.

References

1. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2095-2128. doi:10.1016/S0140-6736(12)61728-0.

2. Mannino DM, Doherty DE, Buist a. S. Global Initiative on Obstructive Lung Disease (GOLD) classification of lung disease and mortality: Findings from the Atherosclerosis Risk in Communities (ARIC) study. *Respir Med*. 2006;100:115-122. doi:10.1016/j.rmed.2005.03.035.

3. Silverman EK. Genetic Epidemiology of COPD. *Chest*. 2002;121(3 Suppl):1S - 6S. http://www.ncbi.nlm.nih.gov/pubmed/11893649. Accessed February 26, 2015.

4. Pillai SG, Ge D, Zhu G, et al. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet*. 2009;5(3):e1000421. doi:10.1371/journal.pgen.1000421.

5. Wilk JB, Chen T-H, Gottlieb DJ, et al. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet*. 2009;5(3):e1000429. doi:10.1371/journal.pgen.1000429.

6. Cho MH, Boutaoui N, Klanderman BJ, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet*. 2010;42(3):200-202. doi:10.1038/ng.535.

7. Hancock DB, Eijgelsheim M, Wilk JB, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet*. 2010;42(1):45-52. doi:10.1038/ng.500.

8. Soler Artigas M, Loth DW, Wain L V, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet*. 2011;43(11):1082-1090. doi:10.1038/ng.941.

9. Kiezun A, Garimella K, Do R, et al. Exome sequencing and the genetic basis of complex traits. *Nat Genet*. 2012;44(6):623-630. doi:10.1038/ng.2303.

10. Huyghe JR, Jackson AU, Fogarty MP, et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet*. 2013;45(2):197-201. doi:10.1038/ng.2507.

11. Peloso GM, Auer PL, Bis JC, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet*. 2014;94(2):223-232. doi:10.1016/j.ajhg.2014.01.009.

12. Igartua C, Myers RA, Mathias RA, et al. Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. *Nat Commun*. 2015;6:5965. doi:10.1038/ncomms6965.

13. Wessel J, Chu AY, Willems SM, et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat Commun*. 2015;6:5897. doi:10.1038/ncomms6897.

14. Chen F, Klein AP, Klein BEK, et al. Exome array analysis identifies CAV1/CAV2 as a susceptibility locus for intraocular pressure. *Invest Ophthalmol Vis Sci*. 2015;56(1):544-551. doi:10.1167/iovs.14-15204.

15. Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDgene) study design. *Epidemiology*. 2011;7(1):1-10. doi:10.3109/15412550903499522.Genetic.

16. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med*. 1999;159(1):179-187. doi:10.1164/ajrccm.159.1.9712108.

17. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*. 2010;11(7):459-463. doi:10.1038/nrg2813.

18. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93. doi:10.1016/j.ajhg.2011.05.029.

19. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164. doi:10.1093/nar/gkq603.

20. Pruitt KD, Brown GR, Hiatt SM, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*. 2014;42(Database issue):D756-D763. doi:10.1093/nar/gkt1114.

21. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315. doi:10.1038/ng.2892.

22. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 2005;15(7):901-913. doi:10.1101/gr.3577405.

23. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653.

24. Eber SW, Gonzalez JM, Lux ML, et al. Ankyrin–1 mutations are a major cause of dominant and recessive hereditary spherocytosis. *Nat Genet*. 1996;13(2):214-218. doi:10.1038/ng0696-214.

25. Imboden M, Bouzigon E, Curjuric I, et al. Genome-wide association study of lung function decline in adults with and without asthma. *J Allergy Clin Immunol*. 2012;129(5):1218-1228. doi:10.1016/j.jaci.2012.01.074.

26. Hansel NN, Ruczinski I, Rafaels N, et al. Genome-wide study identifies two loci associated with lung function decline in mild to moderate COPD. *Hum Genet*. 2013;132(1):79-90. doi:10.1007/s00439-012-1219-6.

27. Teufel A, Weinmann A, Galle PR, Lohse AW. In silico characterization of LZTS3, a potential tumor suppressor. *Oncol Rep*. 2005;14(2):547-551. http://www.ncbi.nlm.nih.gov/pubmed/16012743. Accessed April 13, 2015.

Table 3.1
Characteristics of the COPDgene study population with available spirometry data.

	Non-Hispanic Whites	African Americans
N	6562	3182
Mean Age	62.1 (8.8)	54.7 (7.2)
Gender (% Male)	52.3 %	55.9 %
Mean pack-years smoked	47.4 (26.0)	38.3 (21.6)
Mean FEV_1/FVC	0.64 (0.2)	0.72 (0.1)
Mean FEV_1 % Predicted	0.74 (0.3)	0.82 (0.2)

Table 3.2
Single variant associations with spirometry. Putatively functional SNVs with MAF > 0.5% were tested for association with FEV_1/FVC and FEV_1 percent predicted in 6,051 NHW and 3,321 AA COPDgene subjects.

Non-Hispanic Whites

SNP	Chr	Hg19 Position	Outcome	Beta	P-value	Gene	Function	MAF	P-value AA	MAF AA
rs2070600	6	32151443	FEV_1/FVC	0.04	$1.67x10^{-8}$	AGER	nonsynonomous	0.042	0.920	0.009
rs16969968	15	78882925	FEV_1/FVC	-0.02	$6.03x10^{-10}$	CHRNA5	nonsynonomous	0.366	0.082	0.059
rs16969968	15	78882925	FEV_1 % P	-2.76	$2.26x10^{-10}$	CHRNA5	nonsynonomous	0.366	0.064	0.059

African Americans

SNP	Chr	Hg19 Position	Outcome	Beta	P-value	Gene	Function	MAF	P-value NHW	MAF NHW
rs34664882	8	41543675	FEV_1 % P	-11.59	$2.42x10^{-7}$	ANK1	nonsynonomous	0.017	0.929	0.029

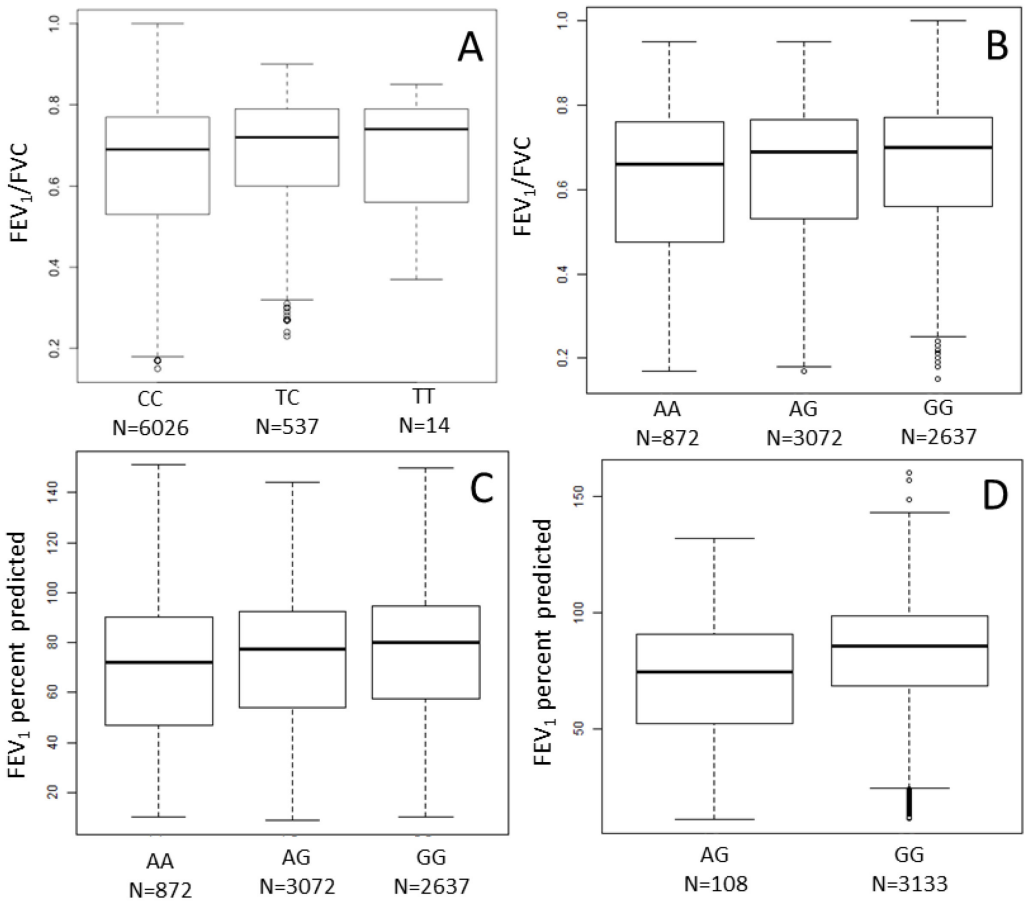


Figure 3.1
Boxplot of spirometry by A) rs2070600 genotype in the *AGER* gene; B) rs16969968 genotype in the *CHRNA5* gene; C) rs16969968 genotype in the *CHRNA5* gene; and D) rs34664882 genotype in the *ANK1* gene.

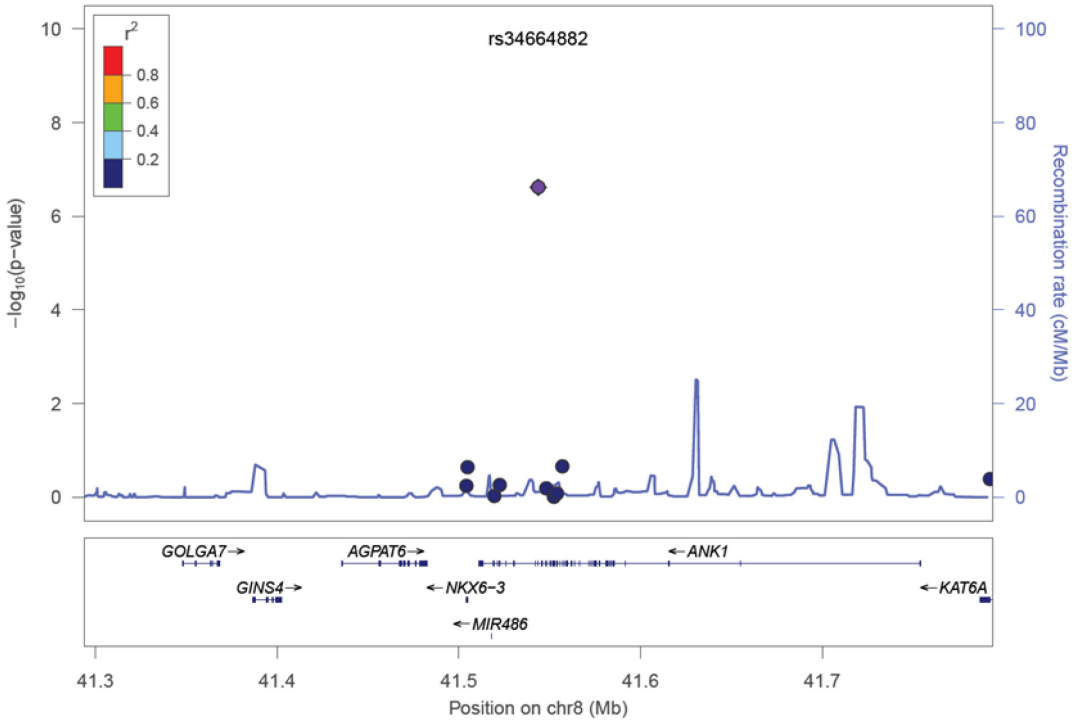


Figure 3.2
Regional association results (- $\log_{10}$ p-value) for SNVs in the *ANK1* gene for FEV_1 percent predicted in AA COPDgene subjects. Linkage disequilibrium (r^2) values from the 1000G African population.

Table 3.3
Genes significantly associated with spirometry. Gene-based analyses were conducted using SKAT. Only rare (MAF < 5%), nonsynonymous, stop-gain, stop-loss, splicing, and non-coding RNA variants in genes with ≥ 5 variants were considered. Genes presented in the table are significant after Bonferroni correction for the number of genes tested ($\alpha = 5.83 \times 10^{-6}$ in NHWs, $\alpha = 5.49 \times 10^{-6}$ in AAs).

Population	Outcome	Gene	Chr	Range	Number of variants	SKAT P-Value
NHW	FEV1/FVC	AGER	6	32148744-32152099	11	2.01×10^{-9}
NHW	FEV1 % P	ProSAPiP1	20	3143272-3149207	7	2.63×10^{-6}
AA	FEV1 % P	ANK1	8	41510743-41754280	19	5.27×10^{-6}

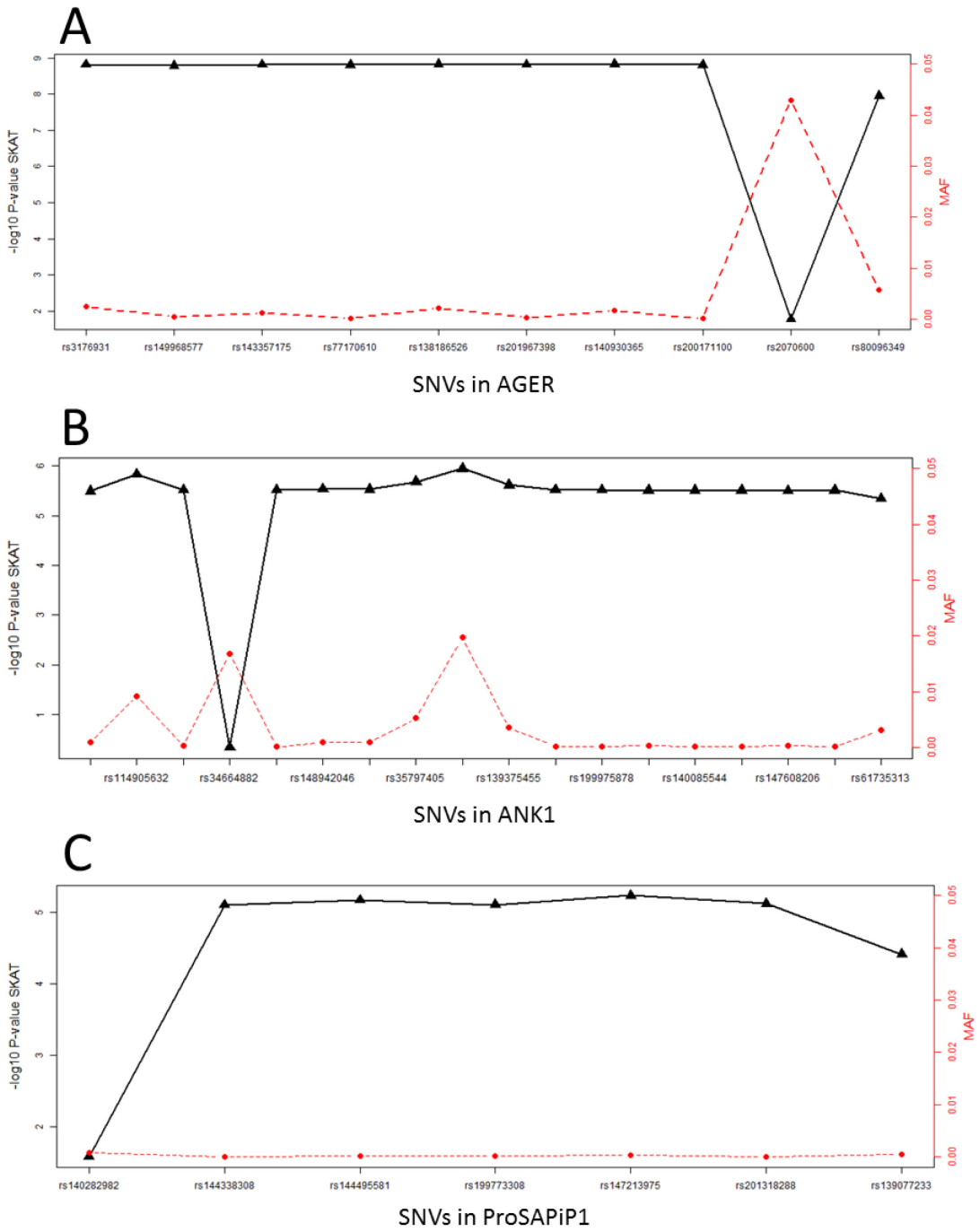
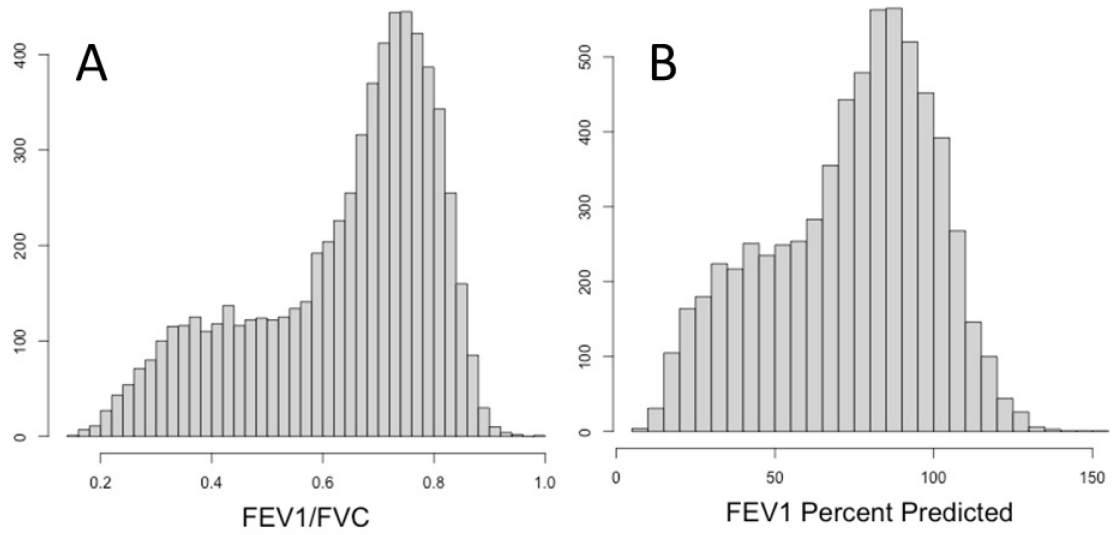
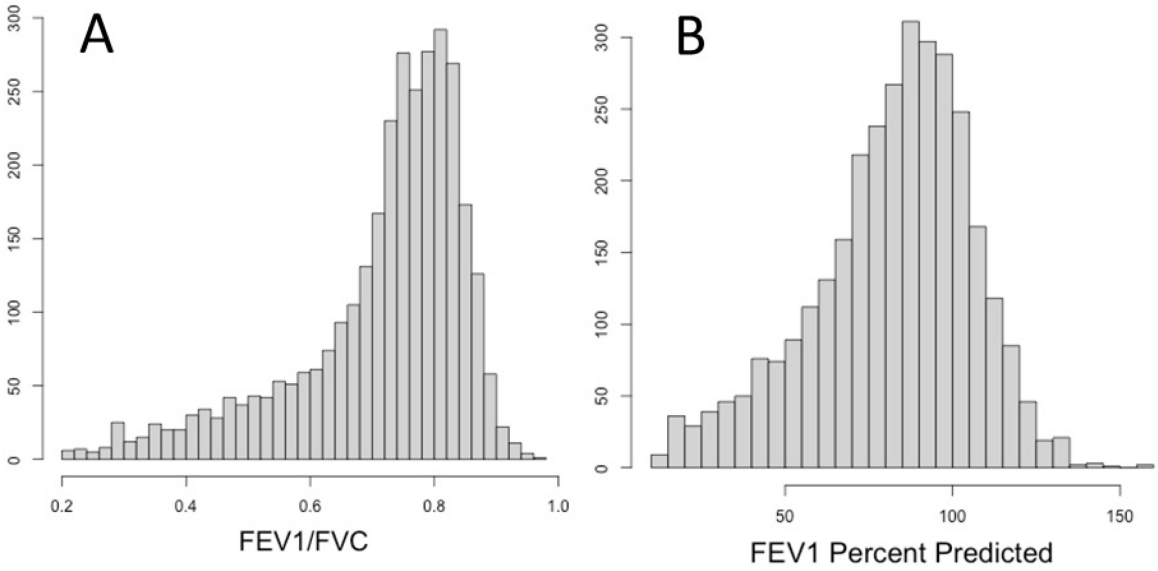


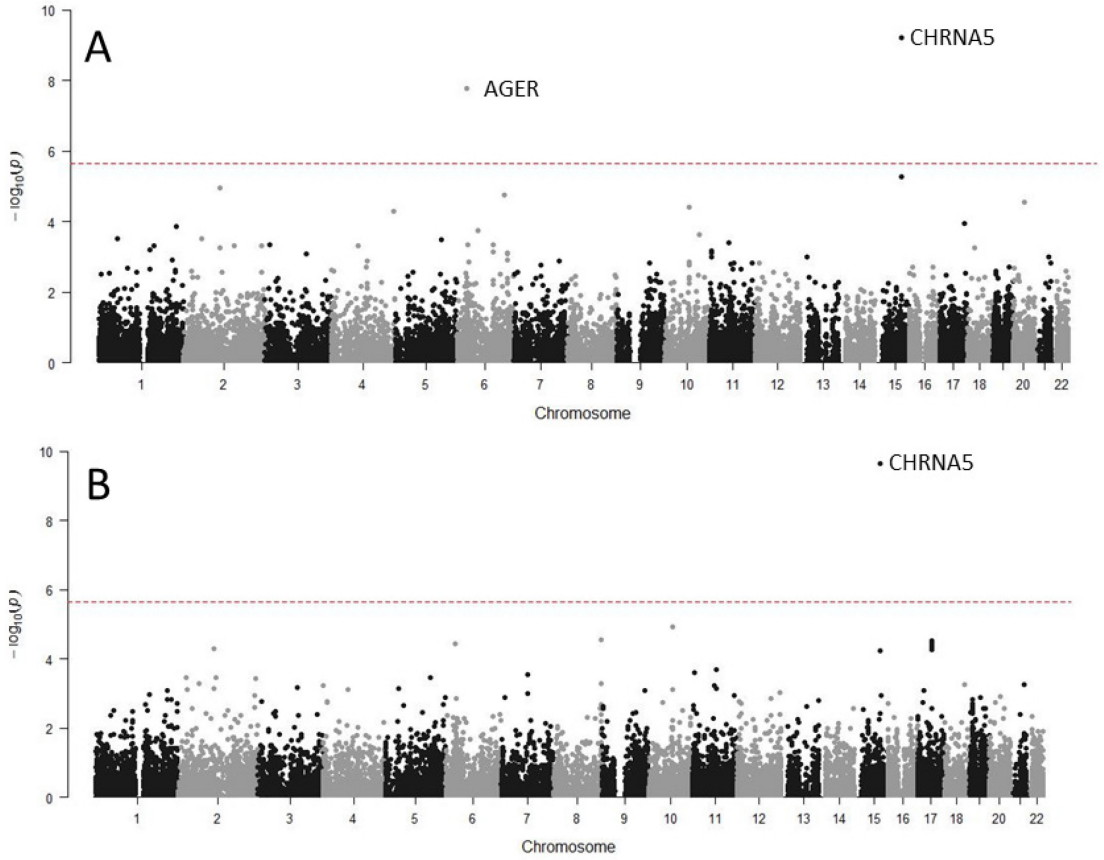
Figure 3.3
Assessment of the relative contribution of each SNV to the gene-based SKAT test for A) SNVs in *AGER*; b) SNVs in *ANK1*; and C) SNVs in *ProSAPiP1*. We re-calculated SKAT p-values removing one SNV at a time, adjusting for age, gender, pack-years smoked, and principal components. Variant MAFs are presented in red.



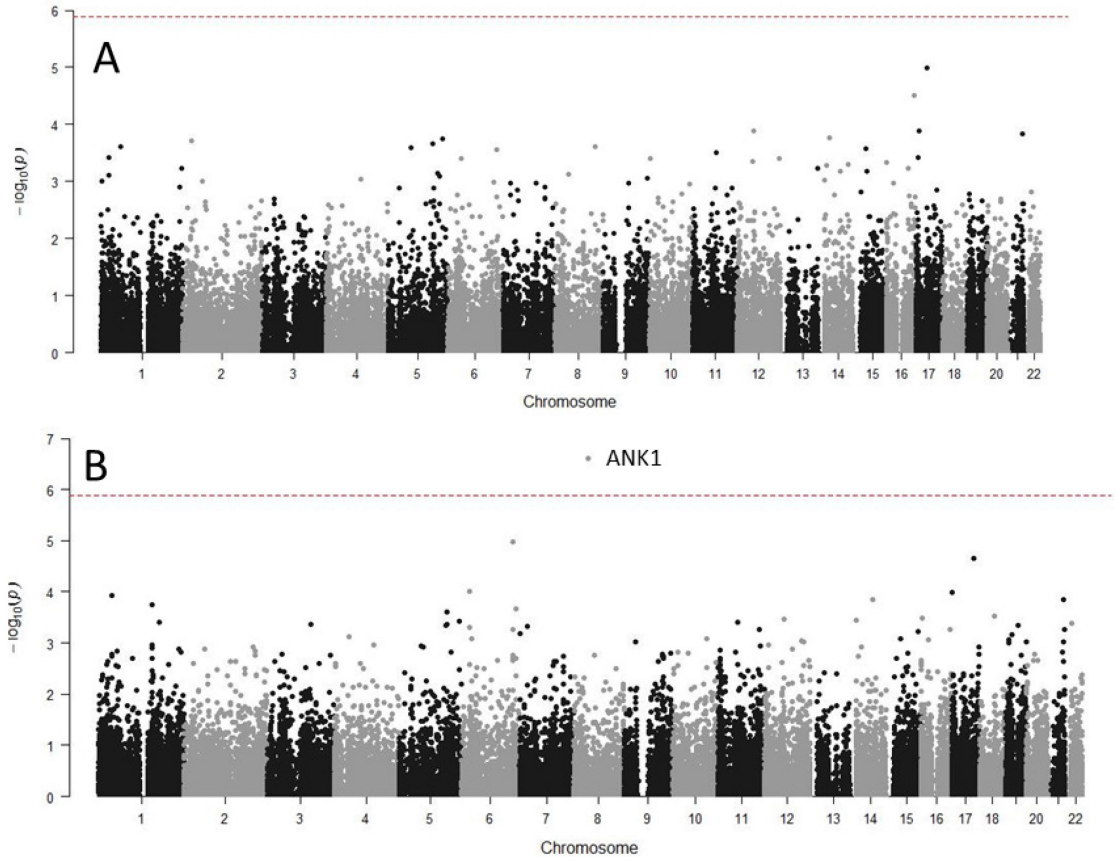
Supplementary Figure 3.1
Histograms of A) FEV_1/FVC and B) FEV_1 percent predicted in NHW COPDgene participants (n=6,562).



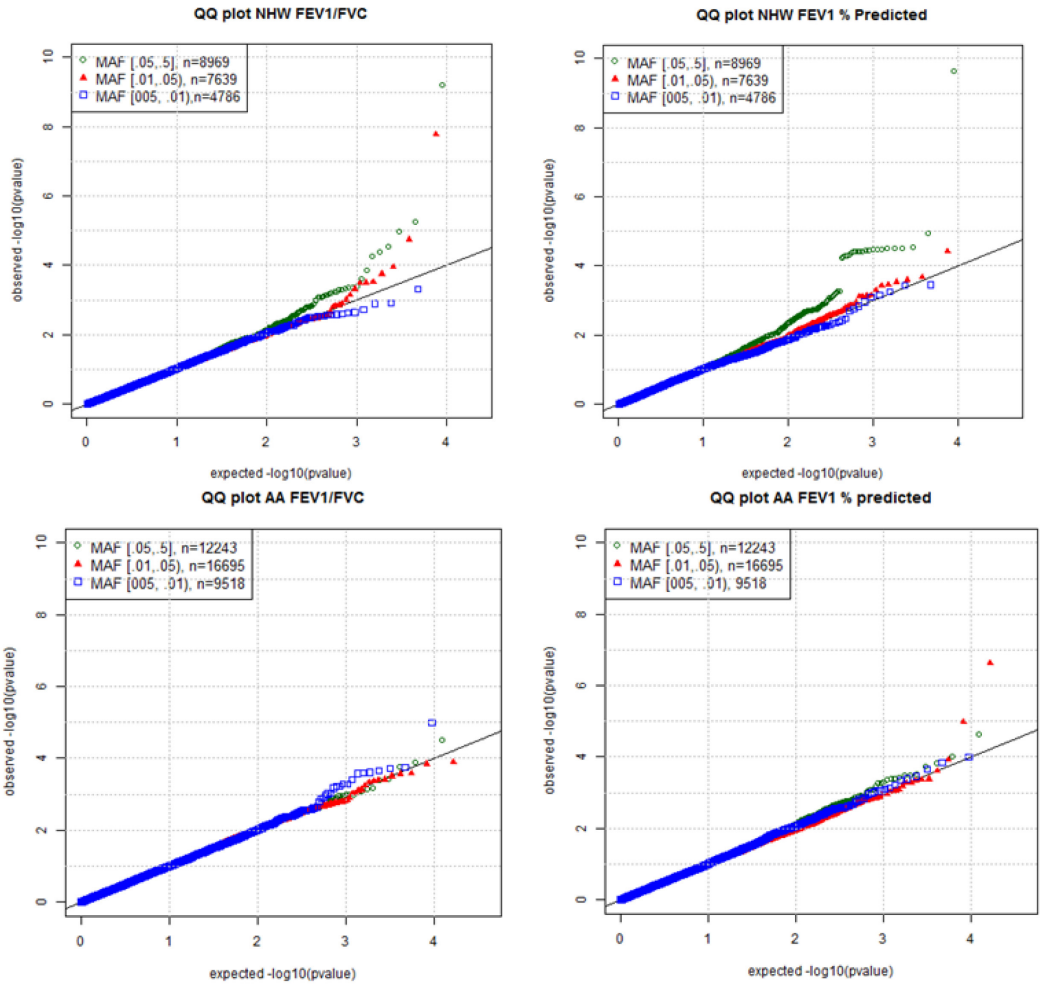
Supplementary Figure 3.2
Histograms of A) FEV_1/FVC and B) FEV_1 percent predicted in AA COPDgene participants (n=3,182).



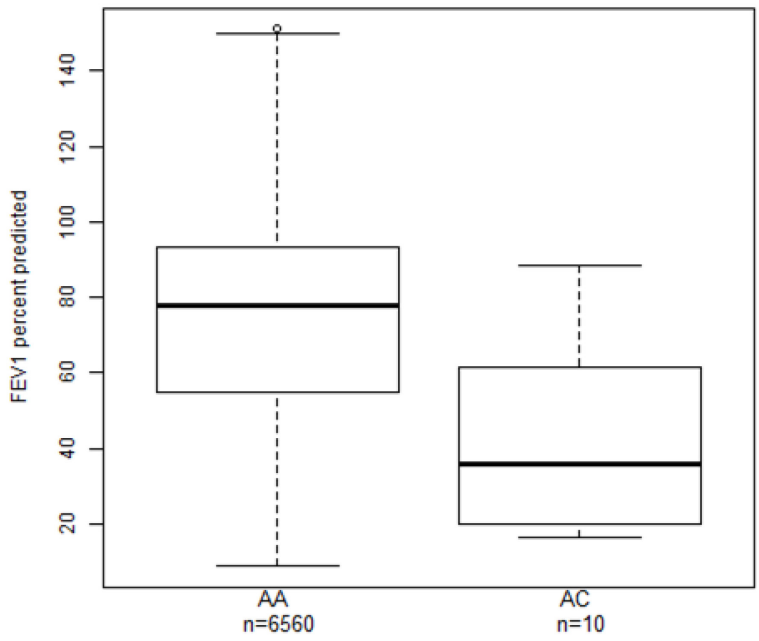
Supplementary Figure 3.3
Manhattan plot of $-\log_{10}$ p-values from a linear regression of outcome on SNV (coded additively) controlling for age, gender, pack-years smoked, and 5 principal components in NHW subjects (n=6,501). Panel A contains results from the FEV_1/FVC analysis. Panel B contains results from the FEV_1 percent predicted analysis. Only functional variants with a MAF > .5% were considered. The red dotted line indicates significance level, as determined by Bonferroni correction for the number of SNVs tested (n=21,394). Significant SNVs are labeled with their gene name.



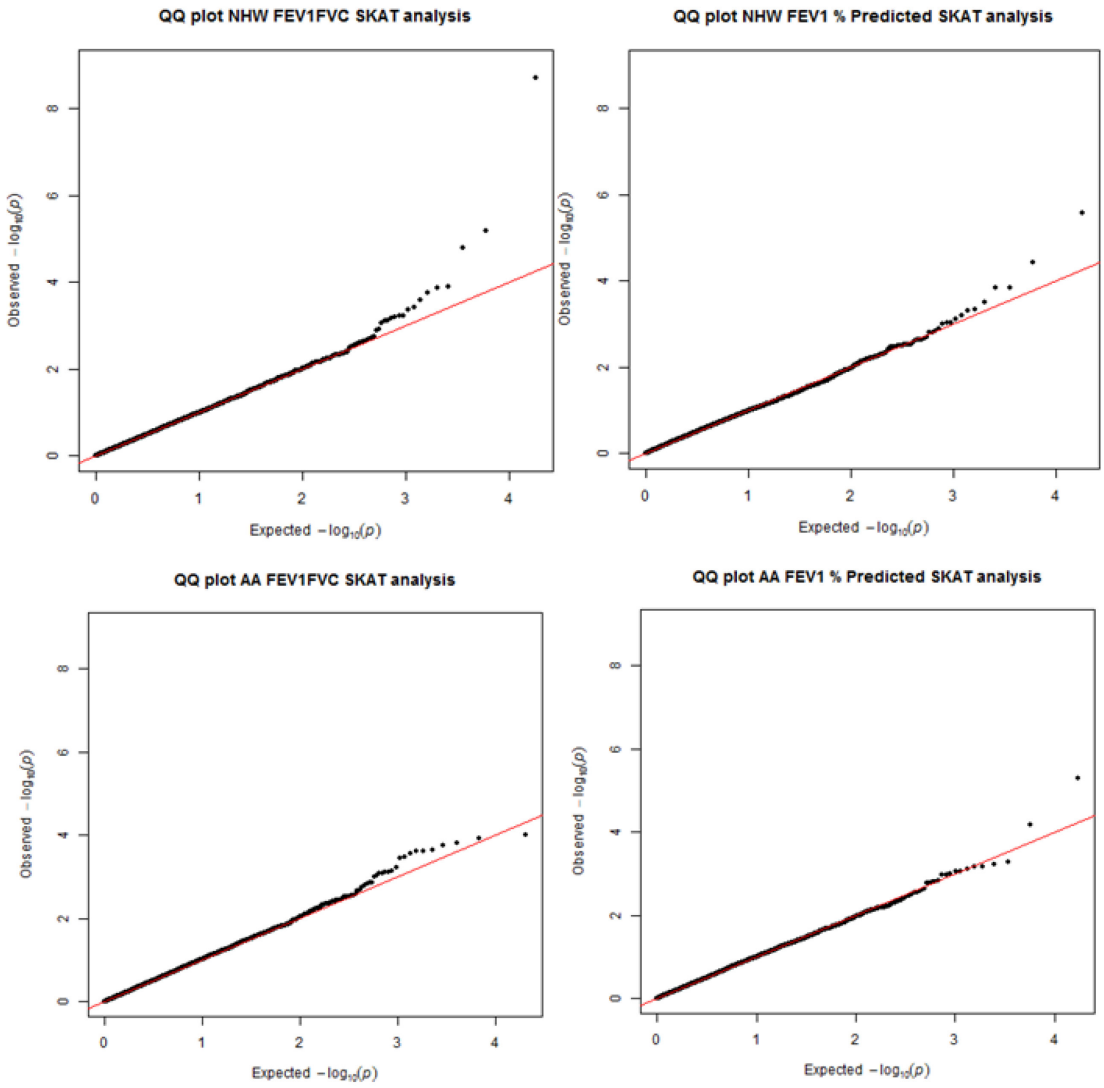
Supplementary Figure 3.4
Manhattan plot of $-\log_{10}$ p-values from a linear regression of outcome on SNV (coded additively) controlling for age, gender, pack-years smoked, and 5 principal components in AA subjects (n=3,221). Panel A contains results from the FEV_1/FVC analysis. Panel B contains results from the FEV_1 percent predicted analysis. Only functional variants with a MAF > .5% were considered. The red dotted line indicates significance level, as determined by Bonferroni correction for the number of SNVs tested (n=38,519). Significant SNVs are labeled with their gene name.



Supplementary Figure 3.5
Quantile-quantile (QQ) plots for single variant analyses in PLINK by minor allele frequency category.



Supplementary Figure 3.6
Boxplot of spirometry by rs140282982 genotype in the *ProSAPiP1* gene. A copy of the minor allele at this SNV reduces FEV_1 percent predicted by -31.7 in NHWs (calculated by linear regression of outcome on genotype, coded additively, controlling for age, gender, pack-years smoked, and the first 5 principal components, p-value = 2.63×10^{-5}).



Supplementary Figure 3.7
Quantile-quantile (QQ) plots for gene-based analysis using SKAT.

Chapter 4. Genetic determinants of decline in lung function in the COPDgene study

Chapter 4. Genetic determinants of decline in lung function in the COPDgene study

Introduction

Spirometric measurements, including the forced expiratory volume in one second (FEV_1) and the forced vital capacity (FVC), reliably reflect the physiological state of the airways, and are used to diagnose chronic obstructive pulmonary disease (COPD)[1,2]. These measures are heritable[3,4] and a multitude of previous research has assessed genetic risk factors for cross-sectional lung function[5–9], including a large scale meta-analysis of 48,201 individuals that identified a total of 26 genetic loci associated with this outcome[9]. However, reduced lung function is likely influenced by both: 1) a failure to attain maximal lung size and function before adulthood; and 2) accelerated lung function decline after maturity. Cross-sectional studies cannot differentiate between these two pathways, and it is probable different risk factors (both genetic and environmental) separately influence each one. However, few studies have assessed risk factors for lung function decline, as this requires longitudinal data.

Fletcher and Peto first established cigarette smoking as a key risk factor for accelerated lung function decline[10], but genetic variation is also hypothesized to play an important role. Heritability estimates of longitudinal change in lung function using family and twin data range from 10% to 39%[4,7,11]. To date, there have been three published genome-wide association studies (GWASs) of lung function decline, all in populations of European ancestry[12–14]. Overall, 6 genes have been identified as associated with lung function decline (*DLEU7*, *TMEM26*, *FOXA1*, *ANK3*, *IL16/STARD5/TMC3*, and *ME3*), none of which overlap with the reported associations from studies of cross-sectional lung function. However, there is little agreement between these 3 published studies regarding

which genes contribute to this decline phenotype, suggesting additional replication and further research is necessary.

To assess the role of genetic risk factors in longitudinal changes in lung function, we used genotyping array data from a large multicenter study of current and former smokers to determine genetic predictors of change in FEV_1 and change in the FEV_1/FVC ratio.

Methods

Study Participants

Subjects who have completed the 5-year follow-up phase of the COPDgene study were included in this preliminary analysis (n=2,000). A complete study protocol for COPDgene has been described elsewhere[15] but briefly, Non-Hispanic Whites (NHWs) and African Americans (AAs) between the ages of 45 and 80 years with a minimum of 10 pack-years smoking history were enrolled at 21 study centers across the United States. In addition to completing detailed questionnaires, pre- and post- bronchodilator spirometry, and volumetric computed tomography (CT) of the chest, participants provided whole blood for DNA extraction and genotyping.

Genotyping and Quality Control (QC)

Samples were genotyped using two array platforms: 1) Illumina's OmniExpress GWAS array, which contains ~730,000 common single nucleotide polymorphisms (SNPs); and 2) Illumina's HumanExome Beadchip (v1.1 or v1.2), which contains ~250,000 functional, primarily low frequency variants in recognized exons of the human genome. Sample and variant quality control (QC) were performed separately for both the genome-wide and exome arrays. A summary of the quality control procedures for the GWAS and exome arrays is provided in Table 1. A complete description of the GWAS array QC is provided at: http://www.copdgene.org/sites/default/files/GWAS_QC_Methodology_20121115.pdf,

and a complete description of the QC for the exome array is provided in Chapter 2 of this dissertation.

After quality control, genotyping data from the genome-wide marker panel and exome arrays were combined (excluding exome array variants also on the genome-wide array), and only variants with a minor allele frequency (MAF) > 0.5% were considered for analysis. This resulted in a total of 1,394 subjects and 654,976 variants available for analysis in the NHW group and 606 subjects and 727,583 variants available for analysis in the AA group.

Principal component analysis

We performed principal component analysis (PCA) separately in NHW and AA subjects using unlinked, polymorphic (MAF > 5%) SNPs that were in Hardy Weinberg equilibrium and not in regions known to have long range linkage disequilibrium (LD)[16]. These inferred principal components summarize genetic ancestry and were applied as covariates in subsequent statistical analyses to control for population stratification within racial group.

Phenotype and covariate assessment

We analyzed: 1) the change in FEV_1 percent predicted (calculated as FEV_1 percent predicted at visit 2 minus FEV_1 percent predicted at visit 1); and 2) the change in FEV_1/FVC ratio (calculated as FEV_1/FVC at visit 2 minus FEV_1/FVC at visit 1). At both baseline and the 5 year follow-up, spirometry measurements were obtained using a standardized spirometer (EasyOne by ndd Medical Technologies) after administering two puffs (180 mcg) of albuterol. FEV_1 percent predicted was calculated using race and gender-specific regression models based on age, age^2, and height^2. We used

demographic data (including age, gender and smoking habits) from the baseline visit for our analysis.

Statistical Analysis

Genome-wide association studies

We performed linear regressions of 5-year change in spirometry on SNP genotype coded additively using PLINK[17] separately in NHWs and AAs. All association models included baseline age, gender, pack-years smoked, baseline spirometry, height, and the first 5 (NHWs) or 6 (AAs) principal components to control for population stratification. SNP-trait analyses using genotype data from the exome array were also adjusted for chip version (v1.1 or v1.2). We declared a variant-trait association to be genome-wide significant if the p-value was less than 5×10^{-8}, and a variant-trait association nominally significant of the p-value was less than $<1 \times 10^{-5}$.

Candidate lung function decline genes

We examined genes previously reported in the literature to be associated with longitudinal changes in lung function to determine if SNPs in or near these genes were associated with change in FEV_1 percent predicted or FEV_1/FVC in the COPDgene cohort. In total, we tested 13 regions, including nine genes/gene clusters with a p-value $< 1 \times 10^{-5}$ in the Tang et al. paper (*ST3GAL3, NFIA, ESRRG/GPATCH2, BAZ2B, TMCO3, IL16/STARD5/TMC3, SV2B, MYH11, CACNG4*)[14], one gene from the Imboden et al. paper (*DLEU7*)[13], and three genes from the Hansel et al. paper (*TMEM26, FOXA1, ANK3*)[12]. Because not all of the sentinel variants in previous studies were available in our data, we annotated all SNPs using ANNOVAR[18] and the RefSeq[19] reference and those annotated to one of the 13 gene regions were assessed for association. This totaled 2,023 variants in the NHW group and 2,193 variants in AA group.

Results

Subject characteristics

Clinical characteristics of subjects with follow-up data are presented in Table 2. This represents 20.3 % of the total NHW cohort and 18.4 % of the total AA cohort expected to complete 5-year follow-up as part of COPDgene. At visit two, 270 subjects (13.5%) had changed smoking status, with 219 subjects (11.0 %) changing from current to former smokers and 51 subjects (2.6%) changing from former to current smokers.

GWAS of change in FEV_1 percent predicted

The mean 5-year change in FEV_1 percent predicted among NHWs was -1.77 ± 9.63, and the mean 5-year change in FEV_1 percent predicted among AAs was -2.97 ± 12.03 (Supplementary Figure 1). There was considerable variability in change in FEV_1 percent predicted (range in NHW: -46.7 to 36.2, range in AA: -39.1 to 48.6) with 820 subjects (41.2 %) increasing FEV_1 percent predicted between visit 1 and visit 2.

Our genome-wide association analyses of change in FEV_1 percent predicted yielded no statistically significant associations (Figure 1). Quantile-quantile (QQ) plots presented in Figure S5 show that there was no evidence of type 1 error inflation (genomic inflation factor NHW analysis = 0.994, genomic inflation factor AA analysis = 0.997). There were 6 SNPs in the NHW analysis and 8 SNPs in the AA analysis that were nominally associated ($p < 1x10^{-5}$) with change in FEV_1 percent predicted (Supplementary Table 1). None of the associated variants overlapped between the NHW and AA analyses.

The most statistically significant SNP among NHWs was rs2867387 ($\beta = 2.28$, $p = 2.7 \times 10^{-6}$) located just downstream of the *FGA* gene. Interestingly, another variant in the 3' UTR of this gene, rs2070022, was also nominally associated with change in FEV_1 percent predicted ($\beta = 2.19$, $p = 6.8 \times 10^{-6}$). The most statistically significant association

in the AA analysis was an exonic variant in the *NUP153* gene (rs16879902, β = -10.9, p = 9.06 x 10^{-7}). This variant has a low minor allele frequency (MAF = 0.03) and results in a missense substitution of an aspartic acid to an asparagine at amino acid position 90.

GWAS of change in FEV_1/FVC ratio

Mean 5-year change in FEV_1/FVC ratio among NHWs was -0.014 ± 0.06, and mean 5-year change among FEV_1/FVC ratio in AAs was -0.021 ± 0.07 (Supplementary Figure 2). Change in FEV_1/FVC ranged from -0.34 to 0.29 in NHWs and from -0.29 to 0.31 in AAs with 684 subjects (37.0 %) increasing their FEV_1/FVC ratio between visit 1 and visit 2.

Our genome-wide association analysis of change in FEV_1/FVC yielded no statistically significant associations (Figure 2). QQ plots presented in Figure S5 show there was no evidence of type 1 error inflation (genomic inflation factor NHW analysis = 1.010, genomic inflation factor AA analysis = 0.999). A total of 11 SNPs in NHWs and 6 SNPs in AAs were nominally associated with change in FEV_1/FVC ratio ($p < 1x10^{-5}$) (Supplementary Table 2). Nominally significant associations did not overlap the FEV_1 percent predicted analysis nor did they overlap between the NHW and AA analyses. The most statistically significant associations in the NHW analysis were two SNPs in high LD (r^2=0.9) located in the intergenic region between the *LOC10050720* and *HNRNPKP3* genes. (rs12574104, β = -0.027, p = 6.25 x 10^{-7} and rs16937161, β = -0.027, p = 9.54 x 10^{-7}). The most statistically significant association in the AA analysis was a variant, rs1998292, in the *RASSF2* gene (β = -0.028, p = 2.24 x 10^{-6}). This variant is located in the intronic regions of the *RASSF2* gene and has a minor allele frequency of 0.13 in AAs.

Candidate lung function decline genes

We sought corroborative evidence of association for genes previously associated with lung function decline (n=13). The most statistically significant SNP association from each of the 13 genes is presented in Table 4 (for change in FEV_1 % predicted) and Table 5 (for change in FEV_1/FVC). Considering $p < 0.001$ as suggestive evidence, none of these SNPs were associated with change in FEV_1/FVC, and three SNPs were associated with change in FEV_1 percent predicted. Interestingly, SNPs within the *NFIA* gene were associated with change in FEV_1 percent predicted in both NHW and AA groups (SNP=rs17377218, p=0.0006 in NHWs and SNP= rs1712138, p=0.0009 in AAs). Additionally, a SNP within the *BAZ2B* gene was also associated with change in FEV_1 percent predicted in NHWs (p=0.0002). This variant is located just downstream of the previously reported variant associated with change in lung function (genomic position = 160223047, genomic position of previously associated = 160250021). These two SNPs are in moderate linkage disequilibrium (r^2 =0.8) in the HAPMAP CEU population[20].

Discussion

Although genetic risk factors for cross-sectional lung function have been well-studied in large scale GWASs[5,9,21], comparatively less information is known about longitudinal changes in spirometry. We sought to assess the role of genetic risk factors in longitudinal changes in FEV_1 percent predicted and FEV_1/FVC using GWAS and exome array data from the COPDgene study, a large multicenter study of current and former smokers.

In our genome-wide analyses, no variants reached statistical significance, suggesting no single variant can explain a large portion of the phenotypic variance in lung function decline. The top SNP in the change in FEV_1 percent predicted analysis, rs17688693, is located in the nucleoporin 153 gene (*NUP153*) on chromosome 6p22. Nucleoporins are

responsible for regulating the movement of macromolecules between the nucleus and cytoplasm, but they have no known role in lung disease[22]. The two top associations in the change in FEV_1/FVC analysis, rs12574104 and rs16937161, were located in the intergenic region between two pseudo-genes (*LOC100507205* and *HNRNPKP3*). The function of these pseudo-genes remains unknown[22].

Additionally, we sought corroborative evidence of association for genes previously associated with lung function decline in three population-based GWASs[12–14]. One gene, nuclear factor 1A (*NFIA)*, showed suggestive evidence of association with change in FEV_1 percent predicted in both the NHW and AA groups (p=0.0006 in NHWs and p=0.0009 in AAs). Genes in the nuclear factor 1 family are highly conserved and function as cellular transcription factors. These genes are known to be essential for lung maturation[23] and variants in this gene have been previously implicated in the asthma plus allergic rhinitis phenotype[24]. Our results support the involvement of this gene in change in lung function among adults.

There are a few limitations to our study. With 1,394 NHW subjects and 606 African American subjects, we have limited power to detect associations, especially those with modest effect sizes. Additionally, we modeled change in lung function by subtracting measures at visit 1 from measures at visit 2 (the “delta method”). Although easy to interpret, this may not fully capture differences in disease trajectory in this population. Specifically, there is evidence those with moderate disease at baseline have a more rapid decline than those with more advanced disease[25–27](Supplementary Figure 4.4 and Supplementary Figure 4.4). Future analyses may benefit from stratification by baseline disease severity. Lastly, changes in lung function are likely to be affected by environmental exposures, especially cigarette smoking. Our analyses controlled for cumulative pack-years smoked, but did not explicitly model changes in smoking status

between visit 1 and visit 2. Additional work is needed to understand how environmental factors and gene by environment interactions affect the rate of lung function decline.

In summary, we performed genome-wide association studies of change in FEV_1 percent predicted and FEV_1/FVC in a large cohort of smokers. This study should be viewed as preliminary analysis, as a sizable portion of COPDgene study participants have yet to complete their 5-year follow-up visit. Additional analyses with this larger sample size are warranted and will provide improved statistical power to detect associations.

References

1. Mannino DM, Doherty DE, Buist a. S. Global Initiative on Obstructive Lung Disease (GOLD) classification of lung disease and mortality: Findings from the Atherosclerosis Risk in Communities (ARIC) study. *Respir Med*. 2006;100:115-122. doi:10.1016/j.rmed.2005.03.035.

2. Vestbo J, Hurd SS, Agustí AG, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease GOLD executive summary. *Am J Respir Crit Care Med*. 2013;187:347-365. doi:10.1164/rccm.201204-0596PP.

3. Wilk JB, Djousse L, Arnett DK, et al. Evidence for major genes influencing pulmonary function in the NHLBI Family Heart Study. *Genet Epidemiol*. 2000;19(July 1999):81-94. doi:10.1002/1098-2272(200007)19:1<81::AID-GEPI6>3.0.CO;2-8.

4. Gottlieb DJ, Wilk JB, Harmon M, et al. Heritability of longitudinal change in lung function. The Framingham study. *Am J Respir Crit Care Med*. 2001;164(9):1655-1659. doi:10.1164/ajrccm.164.9.2010122.

5. Hancock DB, Eijgelsheim M, Wilk JB, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet*. 2010;42(1):45-52. doi:10.1038/ng.500.

6. Pillai SG, Ge D, Zhu G, et al. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet*. 2009;5(3):e1000421. doi:10.1371/journal.pgen.1000421.

7. Wilk JB, Chen T-H, Gottlieb DJ, et al. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet*. 2009;5(3):e1000429. doi:10.1371/journal.pgen.1000429.

8. Cho MH, Boutaoui N, Klanderman BJ, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet*. 2010;42(3):200-202. doi:10.1038/ng.535.

9. Soler Artigas M, Loth DW, Wain L V, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet*. 2011;43(11):1082-1090. doi:10.1038/ng.941.

10. Fletcher C, Peto R. The natural history of chronic airflow obstruction. *Br Med J*. 1977;1(6077):1645-1648. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1607732&tool=pmcentrez&rendertype=abstract. Accessed April 27, 2015.

11. Finkel D, Pedersen NL, Reynolds CA, Berg S, de Faire U, Svartengren M. Genetic and environmental influences on decline in biobehavioral markers of

aging. *Behav Genet*. 2003;33(2):107-123. http://www.ncbi.nlm.nih.gov/pubmed/14574146. Accessed April 27, 2015.

12. Hansel NN, Ruczinski I, Rafaels N, et al. Genome-wide study identifies two loci associated with lung function decline in mild to moderate COPD. *Hum Genet*. 2013;132(1):79-90. doi:10.1007/s00439-012-1219-6.

13. Imboden M, Bouzigon E, Curjuric I, et al. Genome-wide association study of lung function decline in adults with and without asthma. *J Allergy Clin Immunol*. 2012;129(5):1218-1228. doi:10.1016/j.jaci.2012.01.074.

14. Tang W, Kowgier M, Loth DW, et al. Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function. *PLoS One*. 2014;9(7). doi:10.1371/journal.pone.0100776.

15. Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDgene) study design. *Epidemiology*. 2011;7(1):1-10. doi:10.3109/15412550903499522.Genetic.

16. Price AL, Weale ME, Patterson N, et al. Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet*. 2008;83(1):132-135; author reply 135-139. doi:10.1016/j.ajhg.2008.06.005.

17. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1):7. doi:10.1186/s13742-015-0047-8.

18. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164. doi:10.1093/nar/gkq603.

19. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 2007;35(Database issue):D61-D65. doi:10.1093/nar/gkl842.

20. The International HapMap Project. *Nature*. 2003;426(6968):789-796. doi:10.1038/nature02168.

21. Repapi E, Sayers I, Wain L V, et al. Genome-wide association study identifies five loci associated with lung function. *Nat Genet*. 2010;42(1):36-44. doi:10.1038/ng.501.

22. Pruitt KD, Brown GR, Hiatt SM, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*. 2014;42(Database issue):D756-D763. doi:10.1093/nar/gkt1114.

23. Steele-Perkins G, Plachez C, Butz KG, et al. The transcription factor gene Nfib is essential for both lung maturation and brain development. *Mol Cell Biol*. 2005;25(2):685-698. doi:10.1128/MCB.25.2.685-698.2005.

24. Dizier M-H, Margaritte-Jeannin P, Madore A-M, et al. The nuclear factor I/A (NFIA) gene is associated with the asthma plus rhinitis phenotype. *J Allergy Clin Immunol*. 2014;134(3):576-582.e1. doi:10.1016/j.jaci.2013.12.1074.

25. Mohamed Hoesein FAA, Zanen P, Boezen HM, et al. Lung function decline in male heavy smokers relates to baseline airflow obstruction severity. *Chest*. 2012;142(6):1530-1538. doi:10.1378/chest.11-2837.

26. Vestbo J, Edwards LD, Scanlon PD, et al. Changes in forced expiratory volume in 1 second over time in COPD. *N Engl J Med*. 2011;365(13):1184-1192. doi:10.1056/NEJMoa1105482.

27. Sanchez-Salcedo P, Divo M, Casanova C, et al. Disease progression in young patients with COPD: rethinking the Fletcher and Peto model. *Eur Respir J*. 2014;44(2):324-331. doi:10.1183/09031936.00208613.

Table 4.1
Summary of the GWAS and exome array quality control procedures. Genotyping data from the GWAS and exome arrays were combined for analysis, resulting in a total of 1,394 subjects and 654,976 variants available for analysis in the NHWs and 606 subjects and 727,583 variants available for analysis in the AAs. A complete description of the GWAS array QC is provided at: http://www.copdgene.org/sites/default/files/GWAS_QC_Methodology_20121115.pdf, and a complete description of the exome array QC is provided in Chapter 2 of this dissertation.
Definition of abbreviations: HWE= Hardy Weinberg Equilibrium; MAF= minor allele frequency; NHW = Non-Hispanic White; AA = African American; GWAS= genome-wide association study.

GWAS array

Genotyping platform	N variants genotyped	Variant QC filters	Subject QC filters	N subject analyzed	N variants analyzed
Illumina OmniExpress BeadChip	733,200	Call rate < 95% HWE P < 10^{-8} Concordance < 95% MAF < 1%	Call rate < 95% Sex discordance Relatedness Duplicate sample Population outlier	NHW: 1,394 AA: 606	NHW: 630,860 AA: 684,318

Exome arrays

Genotyping platform	N variants genotyped	Variant QC filters	Subject QC filters	N subject analyzed	N variants analyzed
Illumina HumanExome Beadchip v1.1 (NHW only)	247,870	Call rate < 95% HWE P < 10^{-8} Concordance < 95% Freq. difference between chips MAF < 0.5%	Call rate < 95% Sex discordance Relatedness Duplicate sample Population outlier Heterozygosity > /6SD/ GWAS discordance	NHW: 451	NHW: 24,116
Illumina HumanExome Beadchip v1.2	244,770	Call rate < 95% HWE P < 10^{-8} Concordance < 95% Freq. difference between chips MAF < 0.5%	Call rate < 95% Sex discordance Relatedness Duplicate sample Population outlier Heterozygosity > /6SD/ GWAS discordance	NHW: 908 AA: 606	NHW: 24,116 AA: 43,265

Table 4.2
Clinical characteristics of COPDgene subjects with follow up data. Change in FEV_1 percent predicted was calculated as FEV_1 percent predicted at visit 1 – FEV_1 percent predicted at visit 2. Change in FEV_1/FVC ratio was calculated as FEV_1/FVC at visit 1 – FEV_1/FVC at visit 2.

	Non-Hispanic Whites	African Americans
N	1394	606
Mean Age	68.3 (8.3)	59.9 (7.2)
Gender (% Male)	50.5	51.8
Mean pack-years smoked	47.9 (25.2)	39.0 (19.5)
Mean change in FEV_1/FVC	0.014 (0.06)	0.020 (0.07)
Mean change in FEV_1 percent predicted	-1.77 (9.63)	-2.97 (12.03)

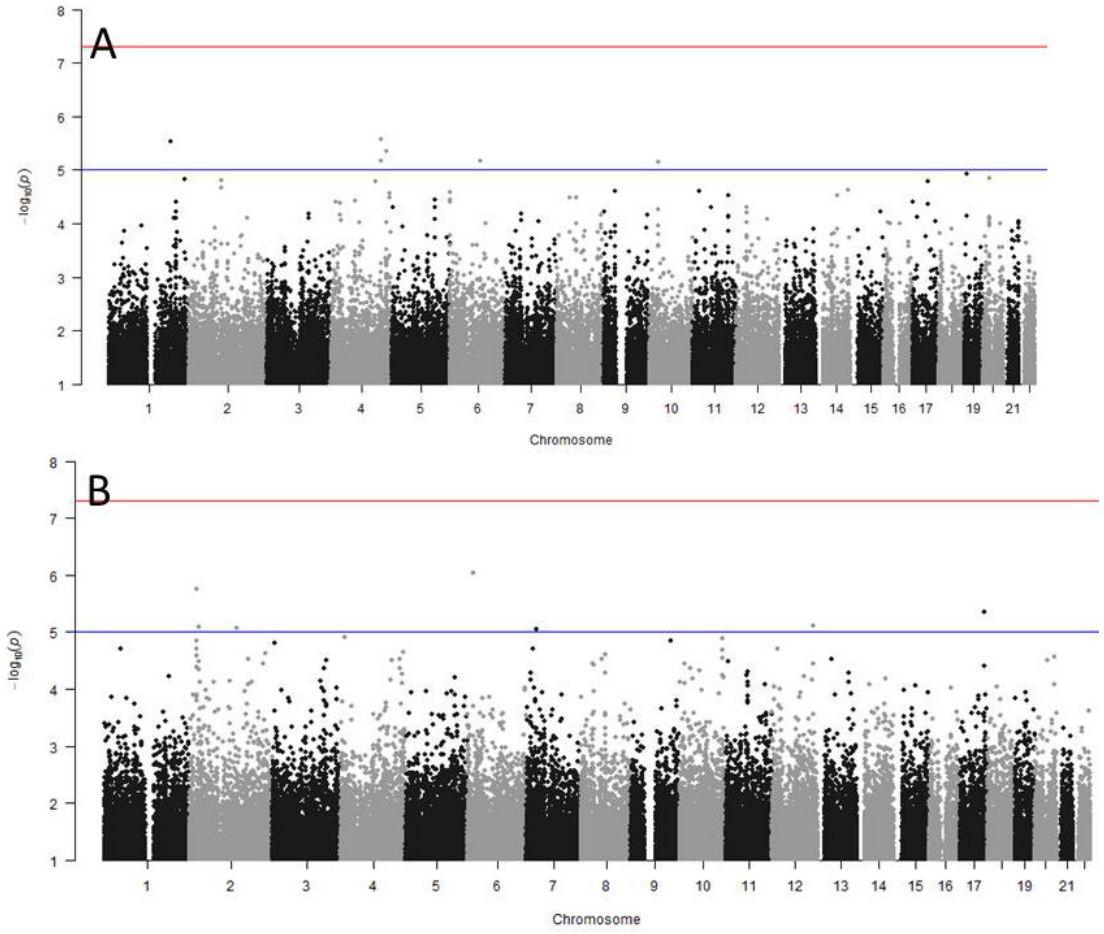


Figure 4.1
Manhattan plot of $-\log_{10}$ p-values from a linear regression of change in FEV_1 percent predicted on SNP (coded additively) controlling for age, gender, pack-years smoked, height, baseline spirometry and principal components. Panel A contains results from the NHW analysis. Panel B contains results from the AA analysis. Only variants with a MAF > .5% were considered. The red dotted line indicates genome-wide significance level ($p < 5 \times 10^{-8}$) and the blue line indicates the nominal significance level ($p < 1 \times 10^{-5}$).

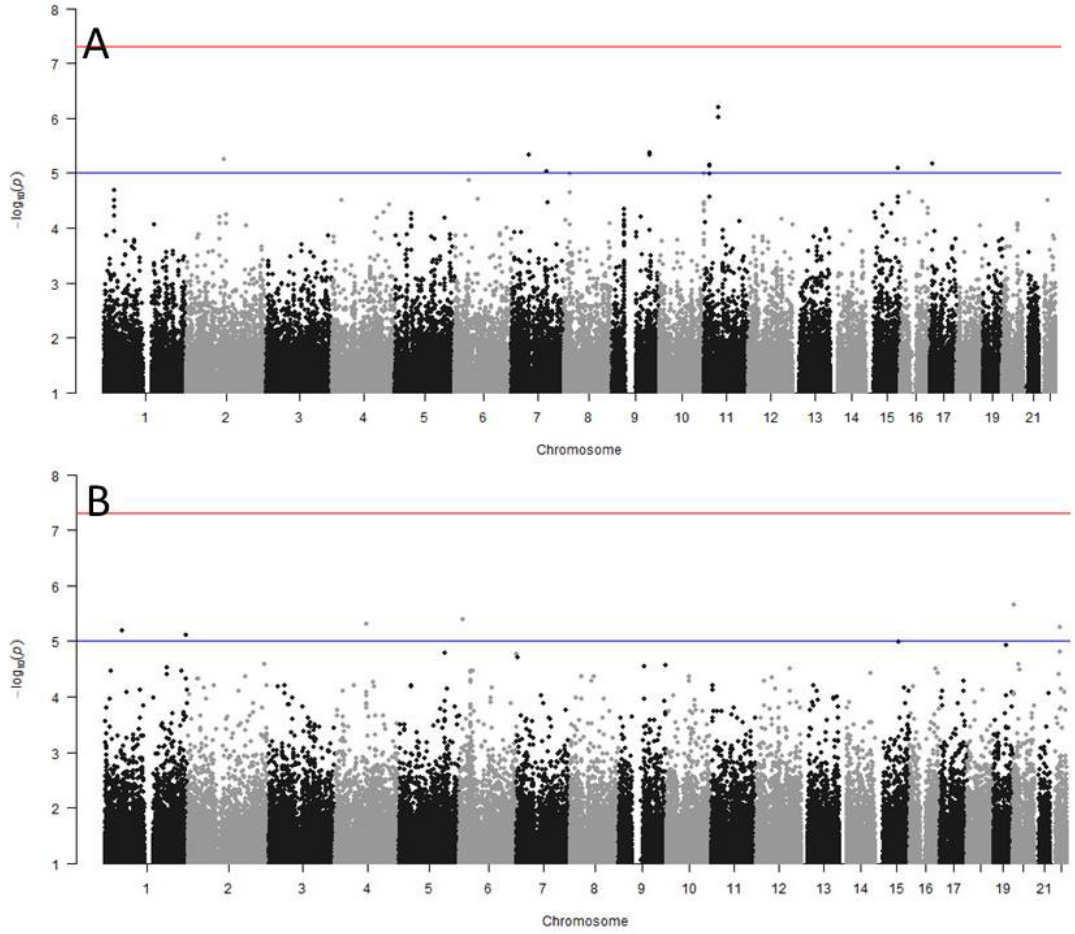


Figure 4.2
Manhattan plot of $-\log_{10}$ p-values from a linear regression of change in FEV_1/FVC on SNP (coded additively) controlling for age, gender, pack-years smoked, height, baseline spirometry and principal components. Panel A contains results from the NHW analysis. Panel B contains results from the AA analysis. Only variants with a MAF > .5% were considered. The red dotted line indicates genome-wide significance level ($p< 5 \times 10^{-8}$) and the blue line indicates the nominal significance level ($p < 1 \times 10^{-5}$).

Table 4.4
Assessment of previously associated lung function decline genes (n=13) with change in FEV_1 percent predicted in COPDgene. All SNPs annotated to the 13 regions were tested for association. The most significant SNP for each region is presented. P-values less than 0.001 are highlighted.

Definition of abbreviations: SNP = single nucleotide polymorphism; P= P-value; NHW=Non-Hispanic White; AA= African American.

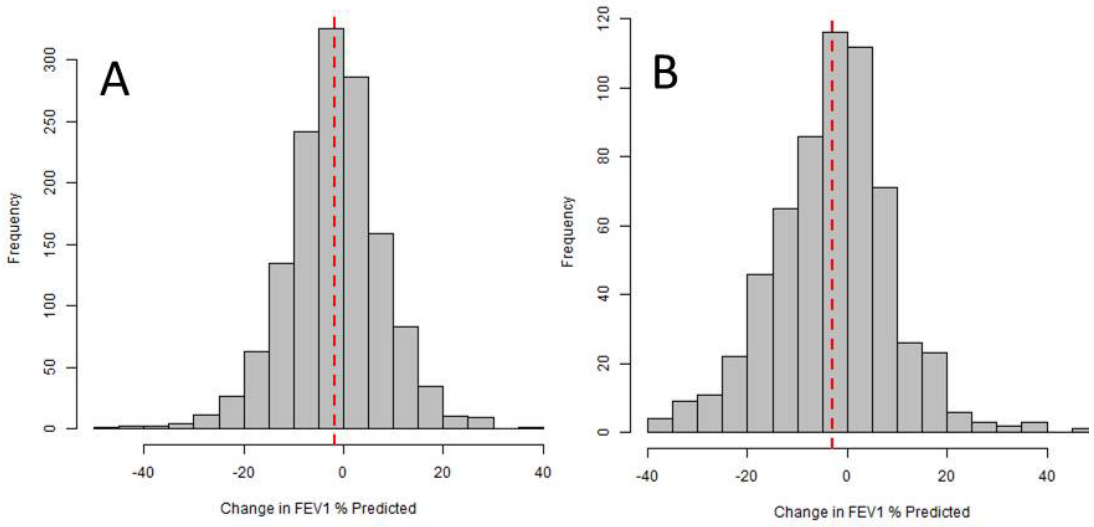
Gene	N SNP NHW	Top SNP NHW	Position NHW	Beta NHW	P NHW	N SNP AA	Top SNP AA	Position AA	Beta AA	P AA
ANK3	305	rs12265962	61787551	5.2280	0.0031	318	61847040	rs11596260	-2.2660	0.0074
BAZ2B	108	rs888629	160223047	-2.5800	**0.0002**	117	160389771	rs13393680	2.6110	0.0266
CACNG4	81	rs2108822	65021998	0.7851	0.0459	86	64970002	rs3826347	2.0200	0.0069
DLEU7	113	rs17074964	51453035	1.8310	0.0016	128	51452751	rs2408213	1.8720	0.0106
ESRRG/GPATCH2	455	rs17045466	217366425	-1.6080	0.0047	507	217365587	rs6669865	-2.3300	0.0053
FOXA1	120	rs1954017	38466498	2.4630	0.0111	133	38212471	rs728088	1.9600	0.0051
IL16/STARD5/TMC3	145	rs4617815	81610902	-1.1420	0.0149	150	81650593	rs56715910	-3.9040	0.0259
MYH11	79	rs8057655	15910474	1.0320	0.0278	79	15950647	rs3851706	-2.0950	0.0484
NFIA	257	rs17377218	61700259	-2.6270	**0.0006**	289	61441952	rs17121389	-4.9490	**0.0009**
ST3GAL3	48	rs3791040	44202733	-0.7776	0.0479	52	44209698	rs16831120	3.0990	0.0433
SV2B	156	rs12902702	91713116	0.9850	0.0175	170	91645528	rs11634420	2.3050	0.0194
TMCO3	42	rs2260722	114188291	0.8866	0.0321	44	114204369	rs1046518	-1.7500	0.0691
TMEM26	114	rs12414729	63421242	-1.4400	0.0641	120	63167404	rs2139780	2.1580	0.0035

Table 4.5

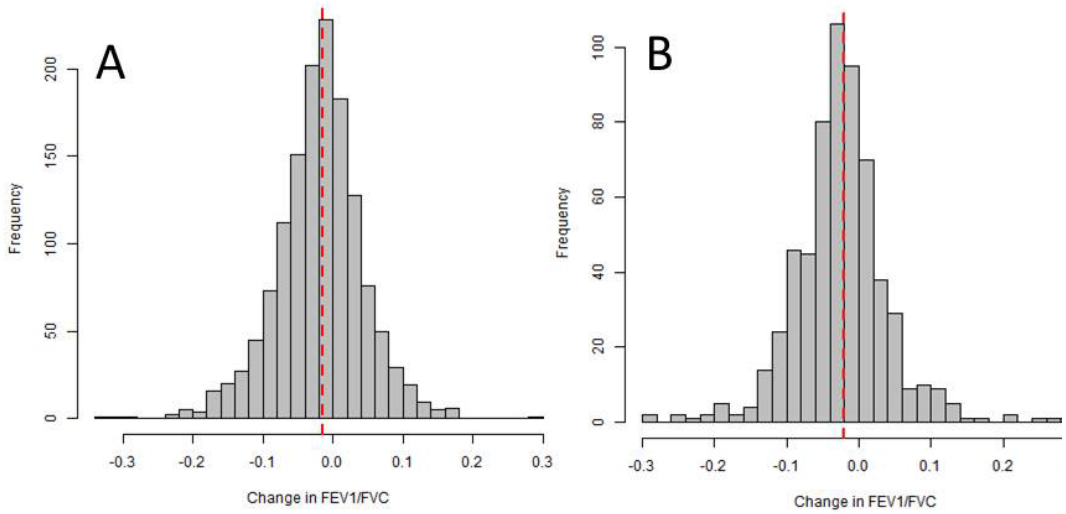
Assessment of previously associated lung function decline genes (n=13) with change in FEV_1/FVC in COPDgene. All SNPs annotated to the 13 regions were tested for association (n=2,023 variants in NHWs, n=2,193 variants in AAs). The most significant SNP for each region is presented. P-values less than 0.001 are highlighted.

Definition of abbreviations: N=number; SNP = single nucleotide polymorphism; P= P-value; NHW=Non-Hispanic White; AA= African American.

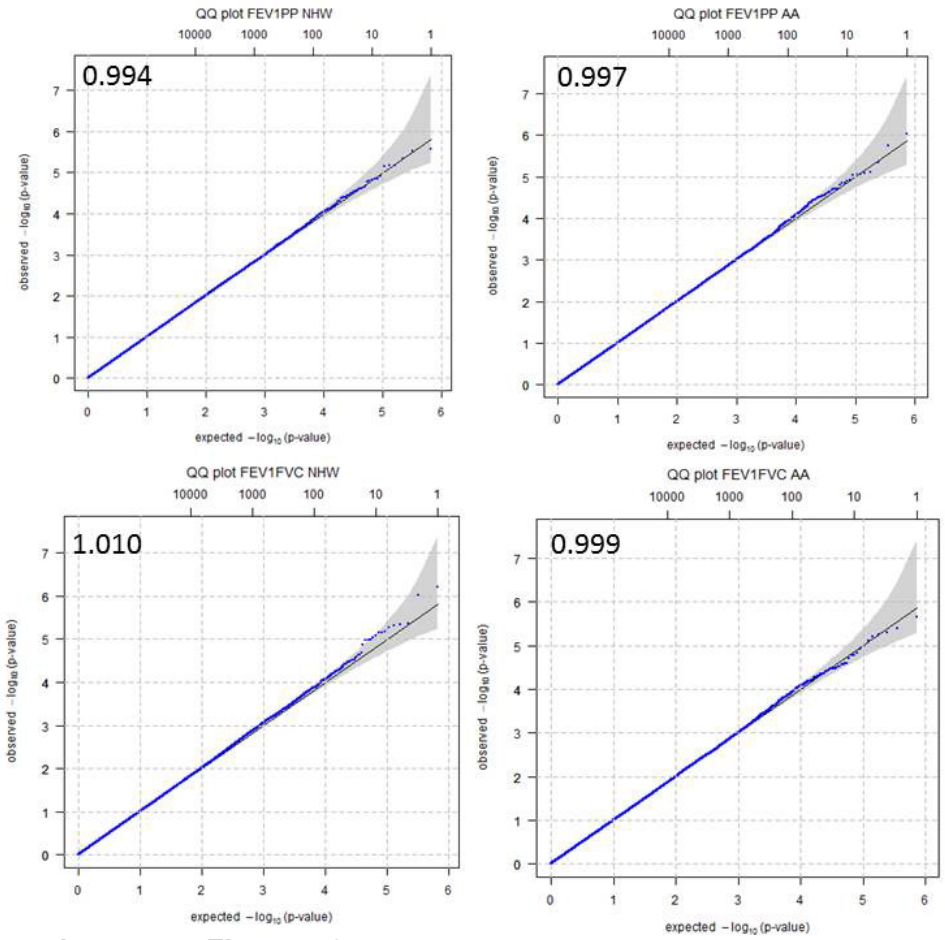
Gene	N SNP NHW	Top SNP NHW	Position NHW	Beta NHW	P NHW	N SNP AA	Top SNP AA	Position AA	Beta AA	P AA
ANK3	305	rs6479700	61963759	-0.0093	0.0218	318	rs2393603	61802342	0.0087	0.0025
BAZ2B	108	rs13393680	160389771	0.0196	0.0022	117	rs888629	160223047	-0.0113	0.0107
CACNG4	81	rs4791017	64882692	0.0185	0.0727	86	rs17644967	64913699	0.0057	0.0506
DLEU7	113	rs706596	51211042	0.0130	0.0019	128	rs706593	51283578	0.0077	0.0066
ESRRG/PATCH2	455	rs1416527	217236801	-0.0122	0.0023	507	rs1335964	216754098	0.0087	0.0074
FOXA1	120	rs177899	38139214	0.0112	0.0078	133	rs1954017	38466498	0.0120	0.0563
Il16/STARD5/TMC3	145	rs7169250	81661639	-0.0120	0.0073	150	rs8034928	81594123	0.0073	0.0089
MYH11	79	rs1050162	15811062	0.0106	0.0125	79	rs12907	15818653	0.0171	0.0672
NFIA	257	rs600411	61355371	-0.0144	0.0065	289	rs1447184	61355010	-0.0070	0.0453
ST3GAL3	48	rs3120803	44386615	0.0059	0.1754	52	rs2108202	44395786	0.0045	0.1188
SV2B	156	rs11073977	91614718	0.0124	0.0020	170	rs10852141	91596017	0.0065	0.0077
TMCO3	42	rs2260335	114175032	0.0116	0.0048	44	rs9805651	114158078	-0.0105	0.0035
TMEM26	114	rs12573098	63163526	-0.0199	0.0063	120	rs7922793	62826952	0.0046	0.0585



Supplementary Figure 4.1
Histograms of change in FEV_1 percent predicted in A) COPDgene NHW subjects completing follow-up (n=1,394) and B) COPDgene AA subjects completing follow-up (n=606). The red dashed line indicates mean change in FEV_1 percent predicted (NHW mean = -1.77, AA mean = -2.97).



Supplementary Figure 4.2
Histograms of change in FEV_1/FVC in A) COPDgene NHW subjects completing follow-up (n=1,394) and B) COPDgene AA subjects completing follow-up (n=606). The red dashed line indicates mean change in FEV_1 percent predicted (NHW mean = -1.77, AA mean = -2.97).



Supplementary Figure 4.3
Quantile-quantile (QQ) plots for single variant analyses of decline phenotypes in PLINK. Genomic inflation factors are presented in the upper left hand corner.

Supplementary Table 4.1
Associations of the most statistically significant SNPs ($P < 1x10^{-5}$) with change in FEV_1 percent predicted in COPDgene NHW (n=1,394) and AA subjects (n=606). Reported results are from a linear regression of change in FEV_1 percent predicted on SNP (coded additively) controlling for age, gender, pack-years smoked, height, baseline spirometry, and principal components.
Definition of abbreviations: CHR = chromosome; SNP = single nucleotide polymorphism; MAF= minor allele frequency.

Non-Hispanic Whites

CHR	SNP	Position	Beta	P-value	MAF	Closest gene(s)	Location
4	rs28673871	155504038	2.277	2.65E-06	0.17	FGA	Downstream
1	rs339597	194914012	2.232	2.93E-06	0.19	LINC01031,KCNT2	Intergenic
4	rs6832095	173935056	1.756	4.48E-06	0.37	GALNTL6	Intronic
6	rs1492964	94031097	-1.994	6.63E-06	0.21	EPHA7	Intronic
4	rs2070022	155504948	2.185	6.78E-06	0.17	FGA	UTR3
10	rs11007747	30078483	1.774	7.08E-06	0.32	SVIL,KIAA1462	Intergenic

African Americans

CHR	SNP	Position	Beta	P-value	MAF	Closest gene(s)	Location
6	rs16879902	17688693	-10.9	9.06E-07	0.03	NUP153	Exonic
2	rs1477472	21798347	5.398	1.75E-06	0.13	APOB,LOC645949	Intergenic
17	rs4432291	70607042	3.442	4.38E-06	0.32	LINC00511	Intronic
12	rs4766921	119350868	4.121	7.79E-06	0.16	SUDS3,SRRM4	Intergenic
2	rs2384298	26156655	-3.459	8.22E-06	0.30	KIF3C	Intronic
2	rs13426918	137035789	4.108	8.60E-06	0.16	CXCR4,THSD7B	Intergenic
2	rs13387850	137035909	4.108	8.60E-06	0.17	CXCR4,THSD7B	Intergenic
7	rs160375	28654701	5.775	9.11E-06	0.08	CREB5	Intronic

Supplementary Table 4.2

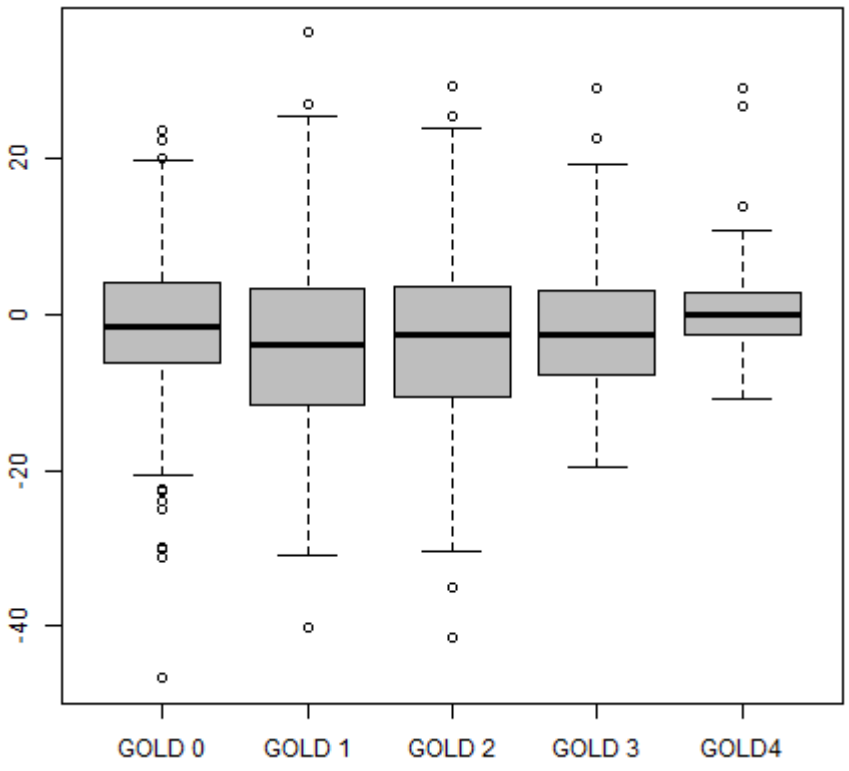
Supplementary Table 2. Associations of the most statistically significant SNPs ($P < 1 \times 10^{-5}$) with change in FEV_1/FVC in COPDgene NHW (n=1,394) and AA subjects (n=606). Reported results are from a linear regression of change in FEV_1/FVC on SNP (coded additively) controlling for age, gender, pack-years smoked, height, baseline spirometry, and principal components.
Definition of abbreviations: CHR = chromosome; SNP = single nucleotide polymorphism; MAF= minor allele frequency.

Non-Hispanic Whites

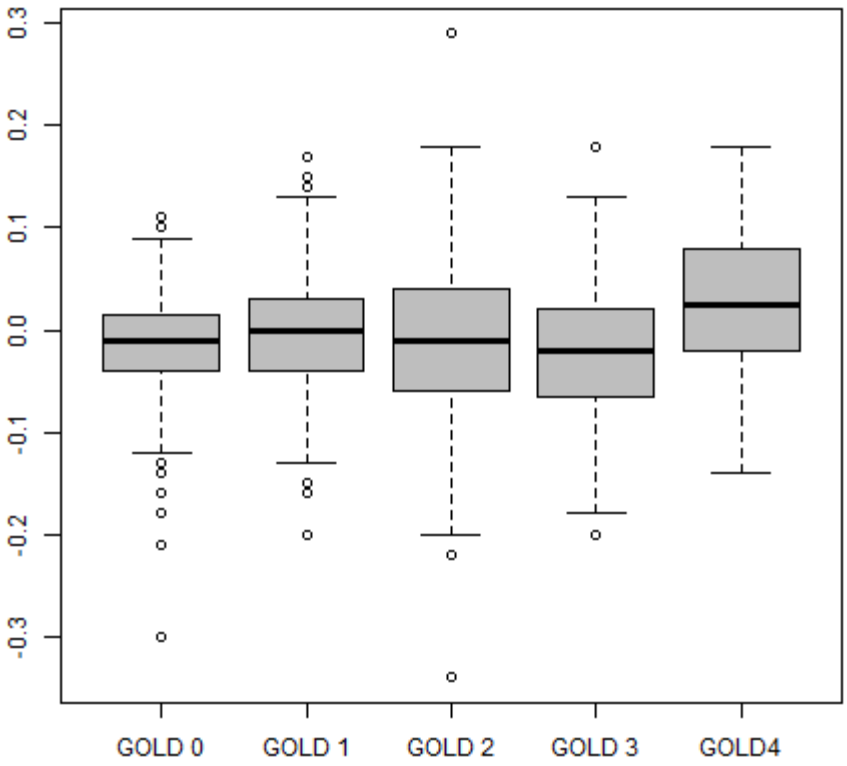
CHR	SNP	Position	Beta	P-value	MAF	Closest gene(s)	Location
11	rs12574104	42922787	-0.027	6.25E-07	0.05	LOC100507205,HNRNPKP3	Intergenic
11	rs16937161	42958117	-0.027	9.54E-07	0.05	LOC100507205,HNRNPKP3	Intergenic
9	rs10980705	113803185	-0.013	4.31E-06	0.23	LPAR1,MIR7702	Intergenic
9	rs7035625	113815456	-0.013	4.59E-06	0.26	LPAR1,MIR7702	Intergenic
7	rs7790298	52218507	-0.017	4.72E-06	0.13	COBL,POM121L12	Intergenic
2	rs13415710	115804950	0.031	5.51E-06	0.03	DPP10	Intronic
17	rs28493751	6441376	-0.017	6.74E-06	0.11	PITPNM3	Exonic
11	rs12365565	18214205	-0.012	7.11E-06	0.29	MRGPRX4,LOC494141	Intergenic
11	rs17567204	18177494	-0.011	7.26E-06	0.34	MRGPRX3,MRGPRX4	Intergenic
15	rs7181753	96844727	-0.013	8.27E-06	0.19	NR2F2	Intronic
7	rs620806	105339137	0.013	9.39E-06	0.18	ATXN7L1	Intronic

African Americans

CHR	SNP	Position	Beta	P-value	MAF	Closest gene(s)	Location
20	rs1998292	4773597	-0.028	2.24E-06	0.13	RASSF2	Intronic
6	rs17340275	7970210	0.074	4.14E-06	0.01	BLOC1S5-TXNDC5	Intronic
4	rs10030500	94406148	0.020	4.95E-06	0.27	GRID2	Intronic
22	rs9625622	29274597	0.026	5.74E-06	0.14	XBP1,ZNRF3	Intergenic
1	rs1288503	53756870	0.023	6.31E-06	0.16	LRP8	Intronic
1	rs6700801	242620466	-0.048	7.61E-06	0.04	PLD5	Intronic



Supplementary Figure 4.4
Change in FEV_1 percent predicted by GOLD classification. GOLD 0 is defined as $FEV_1/FVC > 0.7$. GOLD 1 is defined as $FEV_1/FVC < 0.7$ and $FEV_1 \geq 80\%$ of normal. GOLD 2 is defined as $FEV_1/FVC < 0.7$ and FEV_1 50- 80% of normal. GOLD 3 is defined as $FEV_1/FVC < 0.7$ and FEV_1 30- 50% of normal. GOLD 3 is defined as $FEV_1/FVC < 0.7$ and FEV_1 <30% of normal. GOLD 0 and GOLD 4 subjects have the smallest decline in FEV_1 percent predicted.



Supplementary Figure 4.5
Change in FEV_1/FVC by GOLD classification. GOLD 0 is defined as FEV_1/FVC > 0.7. GOLD 1 is defined as FEV_1/FVC < 0.7 and FEV_1 ≥ 80% of normal. GOLD 2 is defined as FEV_1/FVC < 0.7 and FEV_1 50- 80% of normal. GOLD 3 is defined as FEV_1/FVC < 0.7 and FEV_1 30- 50% of normal. GOLD 3 is defined as FEV_1/FVC < 0.7 and FEV_1 <30% of normal. GOLD 4 subjects have the smallest decline in FEV_1/FVC.

Chapter 5. Summary of key finding and conclusions.

Chapter 5. Summary of key finding and conclusions.

Chronic obstructive pulmonary disease (COPD) is the 3rd leading cause of death worldwide, accounting for approximately 3 million deaths in 2010[1]. Although cigarette smoking is the primary environmental risk factor, COPD susceptibility is related in part to genetic variation[2–5]. Most previous research has focused on identifying associations of common variants with cross-sectional lung function. We hypothesized: 1) rare functional variation also affects lung function; and 2) genetic variation (both common and rare) affects longitudinal changes in lung function. This dissertation addresses these two hypotheses with the following specific aims:

1. To identify functional genetic variants associated with lung function in Non-Hispanic White (NHW) and African American (AA) participants of the COPDgene study using exome array data.
2. To identify genetic variants associated with longitudinal changes in lung function in Non-Hispanic White and African American participants of the COPDgene study using GWAS and exome array data.

Summary of key findings

Exome array analysis of lung function

We aimed to identify rare coding variation associated with quantitative spirometric phenotypes through single variant and gene-based exome array analysis in the COPDgene study. We replicated previously known associations from genome-wide association studies (rs2070600 in *AGER* and rs16969968 in *CHRNA5*), illustrating the potential utility of studying the coding genome to clarify GWAS loci. In addition, we identified two novel associations in the *ANK1* and *ProSAPiP1* genes. Association signals

in these genes were largely driven by a single rare, nonsynonomous variant with a large effect (rs34664882 in *ANK1* and rs140282982 in *ProSAPiP1*). Additional replication and functional validation of these variants is necessary, but together these results suggest searching the exome for additional loci influencing lung function loci may improve our understanding of genetic risk factors for COPD.

Analysis of longitudinal changes in lung function

We assessed the role of genetic risk factors in longitudinal changes in spirometric measures (FEV_1 percent predicted and FEV_1/FVC) using GWAS and exome array data from the COPDgene study. This preliminary genome-wide analysis yielded no statistically significant findings, suggesting no single variant can explain a large portion of the phenotypic variance of lung function changes over a 5-year period. The top SNP associated with change in FEV_1 percent predicted analysis, rs17688693, is located in the nucleoporin 153 (*NUP153*), gene and the top associations in the change in FEV_1/FVC analysis, rs12574104 and rs16937161, were located in the intergenic region between two pseudo genes (*LOC100507205* and *HNRNPKP3*). Replication and functional validation of these associations is necessary. Additionally, we sought confirmatory evidence of association for genes previously associated with decline in lung function[6–8]. Our results support the involvement of one gene, nuclear factor 1A (*NFIA)*, which showed suggestive evidence of association with change in FEV_1 percent predicted in both the NHW and AA groups (p=0.0006 in NHWs and p=0.0009 in AAs).

Although the heritability of lung function decline is established [9,10], genes controlling this phenotype remain largely unknown. To date, large GWAS analyses, including those presented in this dissertation, have failed to produce replicable results[8]. This negative result is likely due to: 1) imprecise measurement of smoking behaviors (especially failure

to account for changes in smoking behavior over time); and 2) differential lung function decline by disease severity (i.e. lung function decline may decelerate in those with more severe disease). However, despite these challenges, identifying risk factors for accelerated disease progression is key to identifying those at highest risk and subsequently improving clinical outcomes. Future large epidemiological studies are warranted.

Strengths and limitations

Strengths

This dissertation has a few key strengths. The COPDgene study is a large, well-characterized study that included both Non-Hispanic White and African American participants. Spirometric measurements were carefully collected, post-bronchodilation, using a standardized protocol. Unlike many population-based cohorts, COPDgene is unique in that many subjects have severe or very severe COPD (FEV_1 percent predicted < 50%).

Additionally, this is one of the first studies to assess the role of rare genetic variation in influencing quantitative measures of lung function and lung function decline. While traditional candidate gene and GWAS analyses can identify indirect associations between tagging SNPs and a quantitative phenotype, the exome array affords the opportunity to directly observe functional variation, potentially aiding to elucidate the mechanisms underlying risk of COPD. Additionally, we queried common and functional variants across the genome in a manner that it is unbiased with respect to previous knowledge of lung function etiology. Therefore, our analyses had the potential reveal causal genes not previously suspected in disease etiology, allowing for hypothesis generation.

Limitations

This dissertation should be interpreted in light of its limitations. While we have detailed which smoking behaviors we analyzed, these common measures of exposure may not fully capture critical aspects of smoking, especially given these data are based on self-report. Specifically, participants may not have reported changes in smoking behaviors over their lifetime, potentially resulting in residual confounding that could bias reported results. Moreover, available measures of smoking behaviors (pack-years smoked and current smoking status) may not completely capture the toxic effects of tobacco.

Additionally, this dissertation used genotype data from Illumina's HumanExome array (chapters 3 and 4) and Illumina's OmniExpress GWAS array (chapter 4). Although these genotyping platforms cover common and coding variants relatively well[11,12], they do not cover all genetic variation. Specifically, rare variants in non-coding regions and very rare variation (e.g. private mutations) are not covered. It is possible that loci affecting lung function will not be represented on the tested platforms, and therefore could not have been identified in this study.

Lastly, our genome-wide analyses of lung function decline had limited sample size (n=1,394 NHWs and n=606 AAs), and therefore limited statistical power to detect genetic associations. Failure to identify statistically significant results likely reflects this limitation. As longitudinal follow-up continues in the COPDgene study, this additional data will greatly improve our ability to detect associations.

Future directions

Additional COPD-related phenotypes

Spirometric phenotypes are reliable indicators of lung function and predict population mortality[13]. However, they do not fully describe the heterogeneous nature of COPD.

Specially, they do not differentiate between small airway disease and emphysema, the two main components of COPD pathogenesis. One unique advantage of the COPDgene study is that all participants underwent computed tomography (CT) imaging, allowing for more refined phenotyping to assess structural lung disease. Future analyses will focus on identifying genetic determinants of COPD-related phenotypes other than spirometry, including, percent emphysema, percent gas trapping, airway wall thickness and chronic bronchitis.

Whole exome and whole genome sequencing

In order to comprehensively study all genetic variation for its association with lung function, sequencing data is necessary. The COPDgene study is generating whole exome and whole genome sequencing on a subset of its 10,000 participants, and these data will allow for more complete evaluation of very rare and non-coding variation with COPD-related outcomes.

Ongoing longitudinal follow-up

The longitudinal analysis included in this dissertation represented approximately 20% of the COPDgene cohort expected to completed 5-year follow-up. The additional sample size afforded as more longitudinal data is collected will aid in the continued study of risk factors (both genetic and environmental) related to changes in lung function over time in this cohort of adults at high risk of COPD.

Public health significance

This study aims to elucidate genetic variants associated with lung function and longitudinal lung function decline. Reduced lung function defines COPD, a common disease with a high global burden[1]. This is one of the first studies to assess the role of rare genetic variants in disease pathogenesis, and we identified novel loci associated

with obstructive pulmonary disease. With replication, identification of associated variants can broaden our understanding of the biological pathways related to this disease. Ultimately, the goal of studying COPD genetics is to provide evidence for clinical practice and prevention. Improved understanding will aid in tailoring treatments to defined COPD subtypes, discovering novel therapeutic interventions and developing effective prevention strategies. This is especially timely given the large and growing burden of COPD[1].

References

1. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2095-2128. doi:10.1016/S0140-6736(12)61728-0.

2. Larson RK, Barman ML. The familial occurrence of chronic obstructive pulmonary disease. *Ann Intern Med*. 1965;63(6):1001-1008. http://www.ncbi.nlm.nih.gov/pubmed/5844558. Accessed March 5, 2015.

3. Lebowitz MD, Knudson RJ, Burrows B. Family aggregation of pulmonary function measurements. *Am Rev Respir Dis*. 1984;129(1):8-11. http://www.ncbi.nlm.nih.gov/pubmed/6703487. Accessed March 5, 2015.

4. Silverman EK, Mosley JD, Palmer LJ, et al. Genome-wide linkage analysis of severe, early-onset chronic obstructive pulmonary disease: airflow obstruction and chronic bronchitis phenotypes. *Hum Mol Genet*. 2002;11(6):623-632. http://www.ncbi.nlm.nih.gov/pubmed/11912177. Accessed February 16, 2015.

5. Hersh CP, Hokanson JE, Lynch D a., et al. Family history is a risk factor for COPD. *Chest*. 2011;140:343-350. doi:10.1378/chest.10-2761.

6. Imboden M, Bouzigon E, Curjuric I, et al. Genome-wide association study of lung function decline in adults with and without asthma. *J Allergy Clin Immunol*. 2012;129(5):1218-1228. doi:10.1016/j.jaci.2012.01.074.

7. Hansel NN, Ruczinski I, Rafaels N, et al. Genome-wide study identifies two loci associated with lung function decline in mild to moderate COPD. *Hum Genet*. 2013;132(1):79-90. doi:10.1007/s00439-012-1219-6.

8. Tang W, Kowgier M, Loth DW, et al. Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function. *PLoS One*. 2014;9(7). doi:10.1371/journal.pone.0100776.

9. Gottlieb DJ, Wilk JB, Harmon M, et al. Heritability of longitudinal change in lung function. The Framingham study. *Am J Respir Crit Care Med*. 2001;164(9):1655-1659. doi:10.1164/ajrccm.164.9.2010122.

10. Wilk JB, Chen T-H, Gottlieb DJ, et al. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet*. 2009;5(3):e1000429. doi:10.1371/journal.pgen.1000429.

11. Grove ML, Yu B, Cochran BJ, et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. *PLoS One*. 2013;8(7):e68095. doi:10.1371/journal.pone.0068095.

12. Peloso GM, Auer PL, Bis JC, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet*. 2014;94(2):223-232. doi:10.1016/j.ajhg.2014.01.009.

13. Mannino DM, Doherty DE, Buist a. S. Global Initiative on Obstructive Lung Disease (GOLD) classification of lung disease and mortality: Findings from the Atherosclerosis Risk in Communities (ARIC) study. *Respir Med*. 2006;100:115-122. doi:10.1016/j.rmed.2005.03.035.

Curriculum Vitae

Margaret Parker

Curriculum Vitae

PERSONAL DATA:
Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health
615 North Wolfe Street W6517, Baltimore, MD 21205
Phone: 617-312-1191
megparker@jhu.edu

EDUCATION:

PhD in Genetic Epidemiology — Expected 2015
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Advisor: Dr. Terri Beaty
Thesis Committee: Dr. Ingo Ruczinski, Dr. Rasika Mathias

MHS in Genetic Epidemiology/Human Genetics — May 2011
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Advisor: Dr. Terri Beaty
Thesis Committee: Dr. Rasika Mathias, Dr. W.H. Linda Kao

BA in Biology, BA in Community Health — May 2007
Tufts University, Medford, MA

PROFESSIONAL EXPERIENCE:

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD — Dec 2009-Present
Senior Research Assistant
Principal Investigator: Dr. Terri Beaty

Tufts Medical Center, Boston, MA — May 2007-July 2009
Clinical Research Coordinator (Feb 2008-July 2009)
Laboratory Technician (May 2007-Feb 2008)
Principal Investigator: Dr. Johanna Seddon

PUBLICATIONS:

Published Peer-Reviewed Articles:

1. Younkin SG, Scharpf RB, Schwender H, **Parker MM**, Scott AF, Marazita ML, Beaty TH, Ruczinski (2014). A genome-wide study of inherited deletions identifies two regions associated with non-syndromic isolated oral clefts. Birth Defects Research.103(4):276-8.

2. Leslie E, Taub M, Liu H, Steinberg KM, Koboldt D, Zhang Q, Carlson J, Hetmanski J, Wang H, Larson D, Fulton R, Kousa Y, Fahkouri W, Naji A, Ruczinski I, Begum F, **Parker MM,** Busch T, Standley J, Rigdon J, Hecht J, Scott A, Wehby G, Christensen K, Czeizel A, Deleyiannis F, Schutte B, Wilson R, Cornell R, Lidral A, Weinstock G, Beaty TH, Marazita M, Murray J (2014). Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci. American Journal of Human Genetics.96(3):397-41.

3. Chen Q, Wang H, Schwender H, Zhang T, Hetmanski JB, Chou YH, Ye X, Yeow V, Chong SS, Zhang B, Jabs EW, **Parker MM**, Scott AF, Beaty TH (2014). Joint testing of genotypic and gene-environment interaction identified novel association for BMP4 with non-syndromic CL/P in an Asian population using data from an internal cleft consortium. PLoS One, 9(10):e109038.

4. **Parker MM***, Foreman MG*, Mathias RA, Abel HJ, Hetmanski JB, Gignoux CR, Burchard EG, Borecki IB, Crapo JD, Silverman EK, Beaty TH and the COPDGene Investigators

(2014). *Admixture Mapping Identifies a Quantitative Trait Locus Associated with Lung Function and COPD-Related Phenotypes*. Genetic Epidemiology, 38(7):652-9.

5. Bureau A, **Parker MM**, Ruczinski I, Taub MA, Marazita ML, Murray JC, Mangold E, Noethen MM, Ludwig KU, Bailey-Wilson JE, Cropp CD, Li Q, Szymczak S, Hetmanski JB, Albacha-Hejazi H, Field LL, Doheny KF, Ling H, Scott AF, Beaty TH (2014). Whole exome sequencing of distant relatives drawn from multiplex families identifies novel potentially damaging variants for oral clefts. Genetics 97(3):1039-1044

6. Bureau A, Younkin SG, **Parker MM**, Bailey-Wilson JE, Marazita ML, Murray JC, Albacha-Hejazi H, Beaty TH, Ruczinski I (2014). *Sharing of rare variants by affected relatives: building evidence for causal variants based on exact sharing probabilities*. Bioinformatics 30(15):2189-96.

7. Younkin SG, Scharpf RB, Schwender H, **Parker MM**, Scott AF, Marazita ML, Beaty TH, Ruczinski
(2014). *A genome-wide study of de novo deletions identifies a candidate locus for non-syndromic isolated cleft lip/palate risk*. BMC Genetics 15(1): 24.

8. Wu T, Schwender H, Ruczinski I, Murray JC, Marazita ML, Munger RG, Hetmanski JB, **Parker MM**, Wang P, Murray T, Redett RJ, Fallin MD, Liang KY, Wu-Chou YH, Chong SS, Yeow V, Ye X, Wang H, Huang S, Jabs EW, Shi B, Wilcox AJ, Jee SH, Scott AF, Beaty TH (2014). *Evidence of gene−environment interaction for two genes on chromosome 4 and environmental tobacco smoke in controlling the risk of non-syndromic cleft palate.* PLoS One 9(2): e88088.

9. Beaty TH, Taub MA, Scott AF, Murray JC, Marazita ML, Schwender H, **Parker MM**, Hetmanski JB, Balakrishnan P, Mansilla MA, Mangold E, Ludwig KU, Noethen MM, Rubini M, Elcioglu N, Ruczinski I (2013). *Confirming genes influencing risk to cleft lip with/without cleft palate in a case parent trio study*. Human Genetics 132(7): 771-81

10. Neale BM, Fagerness J, Reynolds R, Sorbin L, **Parker M**, et al. *Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (LIPC). PNAS.* 2010 Apr 20; 107(16): 7395-400.

Under Review:

1. Bu L, Chen Q, Wang H, Zhang T, Hetmanski JB, Schwender H, **Parker MM**, Chou YW, Chong SS, Zhang B, Jabs EW, Scott AF, Beaty TH (2015). *SNPs in FOX2 gene show evidence of association with NSCL/P in an Asian population.*

2. Begum F, Ruczinski I, Li S, Silverman EK, Cho MH, Lynch D, Curran-Everett D, Crapo J, Scharpf RB, **Parker MM**, Hetmanski JB, Beaty TH (2015). *Identification of a DNA copy number deletion affecting total lung capacity among subjects in the COPDgene Study cohort.*

3. Lutz SM, Cho MH, Hersh CP, Castaldi P, McDonald ML, Regan E, Mattheisen M, DeMeo DL, **Parker MM**, Foreman MG, Make BJ, Jensen RL, Casaburi R, Lomas D, Bhatt SP, Bakke P, Gulsvik A, Crapo JD, Beaty TH, Lange C, Hokanson JE, Silverman EK (2014). *A genome-wide association study identifies novel risk loci for spirometric measures among smokers of European and African Ancestry.*

4. Kim Y, Tilley MK, **Parker MM**, Wojcik GL, Maroo A, Klein AP, Duggal P (2013). *Comparison of the accuracy protein prediction methods to classify deleterious and benign human genetic variation.*

In Preparation:

1. Meeks HD, **Parker MM**, Hetmanski JB, Beaty TH, Munger RG. *Genes related to one-carbon metabolism, nutrient interactions, and risk of isolated orofacial clefts in the GENEVA International Consortium.*

POSTERS:

Presenter:

1. **Parker MM**, Ruczinski I, Mathias R, Beaty TH. Empiric evaluation of rare variant association methods and a simulated quantitative outcome in exome chip data**.** Presented at the Genomics of Common Disease Conference. September, 2014. Potomac, MD.

2. **Parker MM**, Taub MA, Hetrick KN, Ling H, Mathias RA, Hetmanski JB, Albacha-Hejazi H, Scott AF, Ruczinski I, Bailey-Wilson JE, Beaty TH. Comparative analysis of whole exome and whole genome sequencing. Presented at the Johns Hopkins Genetics Research Day, February 2014. Baltimore, MD.

3. **Parker MM**. Taub MA, Hetrick KN, Ling H, Mathias RA, Hetmanski JB, Albacha-Hejazi H, Scott AF, Ruczinski I, Bailey-Wilson JE, Beaty TH. Comparative analysis of whole exome and whole genome sequencing. Presented at ASHG, October 22-26, 2013. Boston, MA.

4. **Parker MM**, Mathias RA, Beaty TH, Gignoux CR, Burchard EG, Hetmanski JB, Silverman EK, Crapo JD, Foreman MG and the COPDgene Investigators. Admixture and Association Mapping Identifies Marker in FAM19A2 Associated with FEV_1/FVC in African-Americans. Presented at ASHG, November 6-10, 2012. San Francisco, CA.

5. **Parker MM**, Mathias RA, Beaty TH, Hetmanski JB, Silverman EK, Crapo JD, Foreman MG and the COPDgene Investigators. Admixture and Association Mapping Identifies Marker in FAM19A2 Associated with FEV_1/FVC in African-Americans. Presented at the Genomics of Common Disease Conference. October, 2012. Potomac, MD.

6. Leslie E, Younkin S, Marazita M, Butali A, Lidral A, Scott AF, Ruczinski I, Taub MA, Hetmanski JB, **Parker MM**, Wang H, Larson D, Harris C, Kobolt D, Steinberg K, Weinstock G, Murray JC, Beaty TH. Common markers identified in case-parent trios confirms 8q24 contains a genetic risk factor for cleft lip with/without cleft palate with substantial heterogeneity across populations. Presented at the Society of Epidemiological Research Meeting. June 2013. Boston, MA.

Co-author:

1. Duggal P, Kim Y, Tilley MK, **Parker MM**, Maroo A, Klein AP. Comparison of the accuracy of protein prediction methods to classify deleterious and benign human genetic variation. Presented at ASHG November 6-10, 2012 San Francisco, CA.

2. Beaty TH, Ruczinski I, **Parker MM**, Duggal P, Taub MA, Li Q, Cropp C, Pugh EW, Wu-Chou YH, Bailey-Wilson JE, Marazita ML, Murray JC, Mangold E, Noethen M Ludwig K, Scott AF. Using whole exome sequencing to identify rare causal variants for oral clefts in multiplex families. Presented at the ASHG, November 6-10 2012 San Francisco, CA.

3. Bailey-Wilson JE, **Parker MM**, Szymczak S, Li Q, Cropp CD, Nothen MM, Hetmanski JB, Ling H, Pugh EW, Duggal P, Taub MA, Ruczinski I, Scott AF, Marazita ML, Murray JC, Mangold E, Albacha-Hejazi H, Beaty TH. Using Whole Exome Sequencing to Identify Rare Causal Variants for Oral Celfts in Multiplex Families with a focus on Syrian Families. Presented at IGES 2012. October 18-20 Stevenson, Washington.

4. Syzmczak, S, Ling H, **Parker MM**, Cropp CD, Ruczinski I, Marazita ML, Murray JC, Mangold E, Mothen MM, Hetmanski JB, Pugh EW, Duggal P, Taub MA, Scott AF, Beaty

TH, Bailey-Wilson JE. Quality Control of variants identified in exome sequencing data in a study of oral clefts. Presented at Genomics of Common Disease Conference, October 2012. Potomac, MD.

5. Qing L, Ling H, **Parker MM**, Cropp CD, Ruczinski I, Marazita ML, Murray JC, Mangold E, Mothen MM, Hetmanski JB, Pugh EW, Duggal P, Taub MA, Scott AF, Beaty TH, Bailey-Wilson JE. Runs of Homozygosity: A Fresh Look at the Whole Exome Sequencing Data in Families. Presented at the Genomics of Common Disease Conference. October, 2012. Potomac, MD.

6. Foreman, M, **Parker MM**, Beaty TH, Kumar R, Aldrich M, Everett D, and the COPDGene investigators. Genetic Ancestry is associated with Lung Function and Chronic Obstructive Pulmonary Disease in the Genetic Epidemiology of COPD (COPDGene) Study. Presented at the ATS conference May 17-22, 2013 Philadelphia, PA.

7. Lutz S, Cho MH, Mattheisen M, McDonald M-L, Regan E, Castaldi PJ, Hersh C, DeMeo D, Bowler R, **Parker MM**, Foreman M, Beaty TH, Laird, Lange C, Hokanson JE, Silverman E. A Genome-wide association study of Pulmonary Function Measures in the COPDGene Study. Presented at the American Thoracic Society Meeting. May 17-22, 2013 Philadelphia, PA.

8. Frederiksen B, Lutz S, Cho MH, Castaldi PJ, **Parker MM**, Kinney GL, McDonald M-L, Santorico SA, Ehringer M, DeMeo DL, Foreman MG, Hersh CP, Make BJ, Curran-Everett D, Lange C, Crapo JD, Silverman EK, Beaty TH, Hokanson JE, for the COPDGene Investigators. Genome-Wide Association Study of Nicotine Dependence in COPDgene. Oral presentation at SRNT March 14th, 2013, Boston, MA.

9. Steinberg K, Koboldt D, Leslie E, Younkin S, Zhang, Marazita M, Butali A, Lidral A, Scott A, Ruczinski I, Taub, M. Hetmanski J, **Parker MM**, Wang H, Larson D, Wilson, Beaty T, Murray J, Weinstock G. Deep Sequencing of non-syndromic cleft lip and palate families. Presented at the Biology of Genomes conference. May 7-11, 2013. Cold Spring Harbor, NY.

AWARDS:

Anna Huffstulter Stiles Award, JHSPH Epidemiology Department	May 2011
Marilyn Menkes Award, JHSPH Epidemiology Department	May 2013
2nd place poster, Johns Hopkins Genetics Research Day	Feb. 2014
Honarable Mention poster, Johns Hopkins Genetic Research Day	Mar. 2015

TEACHING EXPERIENCE:

Teaching Assistant, Johns Hopkins Bloomberg School of Public Health

- Introduction to Genetic Epidemiology (2010, 2013)
- Epidemiologic Methods 3(2011)
- Epidemiologic Methods 4 (2011)
- Epidemiologic Methods 1(2012)
- Principles of Genetic Epidemiology 2 (2012)
- Principles of Genetic Epidemiology 3 (2013)
- Doctoral Proposal Development and Critique (2014, 2015)

FUNDING:

Ruth L. Kirschstein National Research Service Award (CVD T32) Sept 2012-August 2014
Grant #: 5T32HL007024-38

SKILLS:

Programming:
R, UNIX, PLINK, EIGENSOFT, GATK, STATA, Microsoft Office

Leadership:
TA training chair (2013-2014), Johns Hopkins Epidemiology Department
Co-president (2012-2013), Johns Hopkins Epidemiology Student Organization
Social chair (2010-2011), Johns Hopkins Epidemiology Student Organization

COURSEWORK:
Epidemiologic Methods (1-4), Principles of Genetics Epidemiology (1-4), Methods in Biostatistics (1-4), Statistical Computing, PERL for Bioinformatics, Statistics for Genomics, Introduction to Clinical Trials, Causal Inference, Molecular Biology, Human Physiology